Stat.BD

APPLIED BIOSTATISTICS & RESEARCH METHODOLOGY • SESSION 3.5

Working with Data in Excel

Every number you have been shown, made by you instead

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

Session 3.5, between Sessions 3 and 4

Three sessions of numbers, none of them yours

- Session 1: a risk difference of 9.5 points, and about 11 people needing the programme
- Session 2: 339 hypertensive people out of 1,810
- Session 3: 18.7% prevalence, and a denominator that changed halfway down a table

Every one of those was read off a slide.

In plain language

You can already say what these numbers mean. Today you find out where they come from, by making them.

What this session is not

Today

Getting a number out of a column. Counting, dividing, choosing a denominator, and reading a five-number summary off a sheet you are looking at.

Not today

Analysis. No test, no interval, no model. Those arrive in their own sessions, and the SPSS labs do them properly.

Common trap

A spreadsheet will compute anything you ask it for, including things that mean nothing. It has no opinion about whether the question was sensible. That part stays yours.

Act I

A sheet is a dataset, not a report

Stat.BD

One row per person, one column per variable

fx

	A	B	C	D	E	
1	**id**	**sex**	**age**	**education_years**	**waist_cm**	one variable
2	1	1	52	5	88.4	
3	2	2	47	1	76.2	
4	3	1	61	0	81.0	one person
5	4	2	39	7	70.1	
6	5	1	55	2	85.6	

In plain language

The unit of observation from Session 2, laid out flat. One row is one participant. Everything else follows from that.

The header row is one row, and it is the top one

What Excel needs

- Row 1: the variable names, one per column
- Row 2 onwards: nothing but data
- No title, no blank spacer row, no totals in the middle

Common trap

A title across the top and units on a second line looks tidy and makes the sheet uncomputable. Every function then starts counting from the wrong place.

In the clinic

A drug chart has a header block and then the doses. Nobody totals the doses by reading the header. A sheet is the same, except the machine cannot tell which part is which unless you put the names on row 1.

Codes, not labels

Enter

`sex`: 1, 2
`smoker`: 0, 1
`bmi_band`: 1, 2, 3

And keep the key in a README sheet, in the same file.

Common trap

Written out, the same answer arrives as *Yes*, *yes*, *YES*, *Y*, *smoker* and *Yes, daily*. Excel counts those as six different answers, because they are six different pieces of text.

In plain language

A code is a category with the spelling taken out of it.

A blank is not a zero, and 999 is a disaster

fx

	A	B	C	D
1	**id**	**age**	**waist_cm**	**smoker**
2	6	52	88.4	1
3	7	47		0
4	8	61	81.0	1
5	9	39	999	0

not measured: leave it empty

not measured: now a 999 cm waist

Common trap

`=AVERAGE()` skips a blank cell and divides by what is left. It does not skip a 999. It adds it in and reports a mean nobody checks.

Five ways the same twenty patients were entered wrongly

What was done	Why nothing can be computed
Two-row header	No function can find the top of a column
`52 yrs`, `88 cm`	A number with a unit beside it is text
`999` for not measured	Included in every mean and every maximum
Eight spellings of yes	Six categories where there are two
Patient names in column A	A consent problem before it is a data problem

In the clinic

Each of these is a habit from paper records, where a human reads every line and works out what was meant. A machine does not.

The messy sheet in workbook 01 is constructed by us. Nothing on it comes from a study.

Khanam's Table 1 is a report, not a dataset

The register

1,810 rows, one per person. You can count anything you like from it, including things nobody has counted yet.

Table 1

25 rows, one per category. Already counted, already divided. You can only re-divide what somebody chose to print.

In plain language

A table is the answer to questions already asked. A dataset is what you ask the next question of.

Khanam R, et al. *BMJ Open* 2019;9:e026722. PMID 31662350.

Act II

Addresses, formulas, and the denominator that moved

Stat.BD

Every cell has an address

fx

	A	B	C	D	E
1	**id**	**sex**	**age**	**education_years**	**waist_cm**
2	1	1	52	5	88.4
3	2	2	47	1	76.2
4	3	1	61	0	81.0

column C, row 3: **C3**

In plain language

Letter first, then number. Always. `C3` is a place, the way a bed number is a place: it says where to look, not what is there.

A formula starts with an equals sign

fx `=C2+C3`

	A	B	C	D
1	**id**	**sex**	**age**	**scratch**
2	1	1	52	99
3	2	2	47	
4	3	1	61	

the cell shows the answer; the bar shows the formula

In plain language

Without the equals sign it is text and it just sits there. With it, the cell stops holding a value and starts holding an instruction.

A range is two addresses and a colon

- `E2:E1811` one column, every patient
- `B2:E2` one patient, four variables
- `B2:E1811` the whole block

Drag with the mouse and Excel writes the range for you. Read what it wrote before pressing Enter.

Common trap

`E1:E1811` includes the header. `E2:E1810` stops one patient short. Both look right and neither is.

In the clinic

Column-wise is a variable across everybody. Row-wise is everybody's variable for one person. Almost everything you will compute is column-wise.

Drag it down and the addresses move with it

fx `=C2/C$27`

	A	B	C	D
1	**level**	**male_n**	**female_n**	**male_pct**
2	35–44	260	357	30.1
3	45–54	259	290	30.0
4	55–64	167	139	19.3
5	65+	178	160	20.6

one formula, typed once, dragged down four rows

In plain language

A reference is written as an address and remembered as a direction: two columns to my left, same row. Move the formula and the direction is what it keeps.

Counts: Khanam 2019, Table 1. PMID 31662350.

The dollar sign holds one part of an address still

Written	What moves when dragged
`C27`	both
`C$27`	the row is pinned
`$C27`	the column is pinned
`$C$27`	neither

F4 cycles through the four while the formula is open. On a Mac, fn+F4.

In plain language

A percentage has a numerator that changes down the column and a denominator that does not. The dollar sign is how you say which is which.

```
=C2/C$27*100
```

Session 3's error, now visible on screen

Without the dollar sign

fx `=C3/C28`

	A	B
1	**level**	**male_pct**
2	35–44	30.1
3	45–54	145.5
4	55–64	#DIV/0!

With it

fx `=C3/C$27`

	A	B
1	**level**	**male_pct**
2	35–44	30.1
3	45–54	30.0
4	55–64	19.3

Common trap

The first row is right in both. That is what makes it dangerous: the check most people do is to look at the top of the column.

Act III

Counting a categorical column

Stat.BD

COUNT counts numbers; COUNTA counts anything

On the whole register	
`=COUNTA(A2:A1811)`	1,810
`=COUNT(E2:E1811)`	1,810
On the 40-row extract	
`=COUNTA(A2:A41)`	40
`=COUNT(E2:E41)`	38

In plain language

COUNTA asks how many cells have something in them. COUNT asks how many hold a number. The gap between them is your missing data.

Practice register: simulated to match Khanam's printed quartiles; not the study's data.

COUNTIF asks a question of a column


```
=COUNTIF(C2:C1811,">=65")
```

People aged 65 or over	338
As a share of 1,810	18.7%
Khanam's printed 65+ share	18.7%

In plain language

Two arguments: where to look, and what would make a cell count. The condition goes in quotation marks.

Register simulated to match the printed quartiles and the 65+ share. Khanam's other printed age bands are not reproduced.

A percentage is a count over a denominator you chose


```
=COUNTIF(H2:H1811,1)/COUNT(H2:H1811)*100
```

Hypertensive, count	339
Denominator	1,810
Prevalence	**18.7%**

In the clinic

339 and 1,810 are the two numbers Session 2 ended on. Now they are two cells, and the division is yours.

339 is 162 men plus 177 women, which is our arithmetic on Khanam Table 2, not a figure the paper prints.

COUNTIFS makes a two-way table


```
=COUNTIFS(B2:B1811,1,G2:G1811,1)
```

	Smoker	Non-smoker	Total
Male	546	318	864
Female	36	910	946
Total	**582**	1,228	**1,810**

In plain language

Pairs of arguments, as many as you like: range, condition, range, condition. Four cells, four formulas, and the margins check themselves.

Khanam Table 1 counts, reproduced by construction in the simulated register.

Row or column percentage: one reference apart

Of all smokers, how many are men?

```
=COUNTIFS(B2:B1811,1,G2:G1811,1)/COUNTIF(G2:G1811,1)*100
```

93.8%

Denominator: the 582 smokers.

Of all men, how many smoke?

```
=COUNTIFS(B2:B1811,1,G2:G1811,1)/COUNTIF(B2:B1811,1)*100
```

63.2%

Denominator: the 864 men.

Common trap

Same numerator, same data, two answers, and only the question decides which one is the right one. One character differs between the two lines above. Khanam prints the second.

93.8% is ours (546/582). 63.2% is Khanam's printed figure.

100.1 is not always an error

Servings per day, fewer than 2	49.1%
2 to 4	38.6%
5 or more	12.4%
`=SUM(H21:H23)`	**100.1%**

In plain language

Three percentages that each rounded up. Nothing is wrong, nothing is missing, and the column still does not sum to 100.

Khanam 2019, Table 1, total column. PMID 31662350.

Act IV

Summarising a numerical column

Stat.BD

Four functions, one column

Formula	What it answers	On `waist_cm`
`=AVERAGE(E2:E1811)`	the balance point	77.3
`=MEDIAN(E2:E1811)`	the middle person	77.0
`=MIN(E2:E1811)`	the smallest	52.2
`=MAX(E2:E1811)`	the largest	107.9

In the clinic

Same range every time. Once a column is laid out properly, changing the question means changing one word.

Simulated register. The quartiles match Khanam's printed figures; the smallest and largest are our choice, not the study's.

Quartiles and percentiles

`=QUARTILE.INC(E2:E1811,1)`	69.7
`=QUARTILE.INC(E2:E1811,2)`	77.0
`=QUARTILE.INC(E2:E1811,3)`	84.8
`=PERCENTILE.INC(E2:E1811,0.9)`	90.3

In plain language

The second quartile is the median. Excel will give it to you twice and under two names, and that is a sign you have understood it rather than a trick.

Common trap

`.INC` and `.EXC` are two conventions for the same idea and disagree in the second decimal. Use `.INC`, and expect SPSS to differ slightly again.

The five-number summary, as one block

fx	=QUARTILE.INC(cohort!E2:E1811,1)

	A	B	C
1	**statistic**	**formula**	**waist_cm**
2	minimum	MIN	52.2
3	Q1	QUARTILE.INC	69.7
4	median	MEDIAN	77.0
5	Q3	QUARTILE.INC	84.8
6	maximum	MAX	107.9

In plain language

Build it once, on its own sheet, pointing at the data sheet. Then it is a tool rather than a calculation you redo.

77.0 centimetres, and where you have seen it

What Khanam printed

Waist circumference, whole sample:
median **77.0**, IQR **69.7 to 84.8**

What you just computed

median **77.0**,
quartiles **69.7** and **84.8**

Common trap

They agree because the file was built to make them agree. That is a check on the simulation, not a finding about Sylhet. A number from this file is never evidence about anybody.

Khanam 2019, Table 1. PMID 31662350. Register simulated to match those three printed figures.

When the mean and the median part company

	waist_cm	education_years
=AVERAGE()	77.3	3.25
=MEDIAN()	77.0	2
Q1 to Q3	69.7 to 84.8	1 to 5

In plain language

Waist: the two agree. Schooling: the mean sits well above the median, because a few people had twelve years and nobody can have fewer than zero.

Simulated to Khanam's printed quartiles for both columns. PMID 31662350.

STDEV.S or STDEV.P

		Whole register (n = 1,810)	40-row extract (n = 38)
`=STDEV.S()`	divides by $n - 1$	10.02	9.73
`=STDEV.P()`	divides by n	10.01	9.60

In plain language

Use `.S`. You have a sample and you are describing a population you did not measure. On 1,810 people the two agree to the second decimal; on forty they do not.

The card to keep beside the keyboard

Counting	
`COUNT`	how many numbers
`COUNTA`	how many cells are filled
`COUNTIF`	how many meet a condition
`COUNTIFS`	how many meet two
`SUM`	the total

Summarising	
`AVERAGE`	the balance point
`MEDIAN`	the middle person
`MIN, MAX`	the two ends
`QUARTILE.INC`	Q_1, Q_2, Q_3
`STDEV.S`	spread, dividing by $n - 1$

In plain language

Ten functions. Everything in Sessions 4 to 8 is one of these, or a picture of one of these.

What the next session does with all of this

Before the next session

1. Rebuild the five-number summary from scratch, on a new sheet
2. Repair the messy sheet into something computable
3. Compute both percentages for smoking by sex, and write down which question each one answers

In plain language

The next session asks what a mean, a median and a spread actually tell you about a column, and when each one lies. You will arrive able to compute all three.

In the clinic

Answers are in `S3-5-05-completed.xlsx`. Open it after trying, not before: the value is in getting it wrong once.

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. doi:10.1136/bmjopen-2018-026722. CC BY-NC: numbers re-typeset as facts, no text 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. Paywalled: public figures only.
3. Chowdhury MAB, Islam M, Rahman J, et al. Changes in prevalence and risk factors of hypertension among adults in Bangladesh. *PLoS One* 2021;16(12):e0259507. PMID 34855768. PMC8638884. doi:10.1371/journal.pone.0259507. CC BY 4.0.

Provenance. The practice register is **simulated**, matched to Khanam's printed quartiles and to the 65+ and hypertension shares; its tails, and every relationship between columns not named in its README, are ours. 339, 93.8%, 9.5 points, 11, 15.3% and every standardised figure are **our arithmetic**, not printed by any of the three papers.