Stat.BD

CLASS NOTES • CHAPTER 3.5

Working with Data in a Spreadsheet

Ten functions, one dollar sign, and the shape a dataset has to be in

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

Mentor: Muhtasim Munif Fahim, MSc

Chapter 3.5

Working with Data in a Spreadsheet

A spreadsheet computes whatever it is asked for, including things that mean nothing. Almost every error that survives to publication began as a denominator the software was never told to hold still.

Chapters 1 to 3 established what a research question is, how a population differs from a sample, and how a categorical column is described. All of that was done by reading numbers other people had computed. This chapter closes the gap: it is about producing those numbers, in the software a clinical researcher actually has, on a dataset laid out the way analysis software requires.

The order is deliberate. Section 1 comes first because nothing else is possible until the shape of a dataset is right, and because a badly shaped sheet is the commonest reason a statistical package refuses to open a file. Section 2 covers cell addresses and the one piece of notation that separates a working percentage column from a broken one. Sections 3 and 4 cover the two kinds of question a descriptive analysis asks, of a categorical column and of a numerical one. Section 5 is a reference table, and section 6 lists the error messages and what each one means.

The spreadsheet used throughout is Microsoft Excel. The function names are those of Excel for Microsoft 365; Google Sheets and LibreOffice Calc accept all of them unchanged, and where a keystroke differs on macOS the difference is given.

What this chapter establishes

1. That a dataset has a required shape, one row per unit of observation and one column per variable, and why every departure from it costs something later.
2. That a missing value recorded as a number is arithmetically invisible and therefore more dangerous than one left blank.
3. That a cell reference is remembered as a direction rather than an address, and that the dollar sign is how a fixed denominator is declared.
4. How to count a categorical column under one condition and under two, and how the choice of denominator turns the same two counts into two different percentages.
5. How to produce a five-number summary, a mean and a standard deviation from a numerical column, and which standard deviation to use.
6. What each of the five common error values means and what causes it.

Notation

n	the number of observations in a sample
x_i	the value of a variable for the ith observation
$\bar{x}$	the sample mean
s	the sample standard deviation, computed with divisor $n - 1$
σ	the population standard deviation, computed with divisor n
a	a count: the number of observations meeting some condition
N	the denominator chosen for a proportion
w_k	the weight given to the kth stratum
p_k	the rate observed in the kth stratum

Contents

What this chapter establishes 1
Notation 2

1 The shape a dataset has to be in 5

1.1 Codes rather than written answers 5
1.2 Missing values 6
1.3 A dataset and a published table are not the same object 6

2 Addresses, ranges, and the dollar sign 6

2.1 Relative and absolute references 7

3 Counting a categorical column 8

3.1 Counting under a condition 8
3.2 The same numerator, two denominators 8
3.3 Percentages that do not sum to 100 9

4 Summarising a numerical column 9

4.1 Centre and the five-number summary 9
4.2 When the mean and the median separate 10
4.3 Standard deviation, and which one 10
4.4 Two quantities built from a mean and a standard deviation 10
4.5 Weighted sums 11

5 The function reference 11

6 Error values, and what each one means 12

7 Provenance of the figures in this chapter 13

8 Chapter summary 13

9 Key terms 14

10 Exercises 16

11 Solutions 18

12 References and further reading **20**

13 For students in the mentorship programme **20**

1 The shape a dataset has to be in

A clinical record and a dataset are different objects with different purposes, and confusing them is the origin of most of the difficulty that follows. A record is written so that a human can read one patient's story; a dataset is written so that a machine can read one variable across all patients. The record can carry a heading, a note in the margin, an abbreviation the ward understands, and a blank that means whatever the reader infers. A dataset can carry none of those, because the machine does not infer.

Unit of observation
The entity one row of a dataset describes, usually a participant.

One rule follows, and it is short. One row holds one unit of observation, and one column holds one variable. A study of 1,810 adults has 1,810 rows and as many columns as there are variables measured. Row 1 holds the variable names and nothing else. Row 2 onwards holds nothing but data: no title, no blank spacer, no subtotal partway down, and no second header line carrying the units.

> **Intuition**
>
> The shape is not a matter of neatness. Every function a spreadsheet provides operates on a rectangle of cells. If the rectangle contains anything that is not data, the function either refuses or, worse, quietly includes it.

1.1 Codes rather than written answers

A categorical variable is recorded as a code: `1` and `2` for sex, `0` and `1` for a yes-or-no variable, `1`, `2`, `3` for an ordered band. The reason is arithmetic rather than tidiness. Written out by hand, the same answer arrives as *Yes*, *yes*, *YES*, *Y*, *smoker* and *Yes, daily*, and a spreadsheet treats those as six distinct values because they are six distinct pieces of text. A count of smokers then returns whichever spelling happened to be asked for.

Codebook
The record of what each code means, kept with the data.

Codes cannot be read, which is the usual objection to them, and they are not meant to be: their meanings live in a codebook, conventionally a separate sheet in the same file. Keeping the two together is what makes the file self-describing three years later, which is when it will actually be needed.

Codes do not make an ordering where none exists. Coding sex as 1 and 2 creates no sense in which 2 exceeds 1, and a mean of that column is a number with no referent. Coding a BMI band as 1, 2, 3 does preserve an order, and that is what allows the cumulative reading Chapter 3 used. It still does not make the step from 1 to 2 the same size as the step from 2 to 3.

> **Watch out**
>
> A spreadsheet will compute the mean of a column of codes without complaint. The mean of a sex column coded 1 and 2 is a real number, correctly calculated, and it means nothing whatsoever.

1.2 Missing values

A value that was not measured is left blank. It is not written as zero, because zero is a measurement, and it is not written as 999 or −1 or `NA`, because a spreadsheet cannot distinguish a sentinel number from a real one.

Why that matters is the asymmetry of the consequences. A blank cell is ignored by every summary function: it is not counted, not summed, and not included in a mean. A 999 is included in all three. Two sentinel values of 999 in a column of 1,810 waist measurements raise the mean by roughly a centimetre and inflate the standard deviation badly. They also produce a maximum of 999, which is the only one of the three anybody is likely to notice.

Definition 1.1 • What each counting function counts

`COUNTA` returns the number of cells in a range that are not empty, whether their contents are numbers or text. `COUNT` returns the number of cells holding a numeric value. For a numerical variable, the difference between the two is the number of missing observations.

Example 1.1 • Counting what is missing

On the 40-row extract of the practice register, `=COUNTA(A2:A41)` returns 40 and `=COUNT(E2:E41)` returns 38. Two waist measurements are absent. `=AVERAGE(E2:E41)` returns 77.6, having divided the total by 38 rather than by 40.

The STROBE reporting guideline asks, at item 14(b), for the number of participants with missing data for each variable of interest. Subtracting `COUNT` from `COUNTA`, column by column, is that item performed in about ten seconds.

1.3 A dataset and a published table are not the same object

A dataset can answer questions nobody has asked yet. A published Table 1 can only be redivided, and only along the divisions its authors chose. The survey of hypertension in rural Sylhet enrolled 1,810 adults; its Table 1 occupies twenty-five rows. From the 1,810 rows it is possible to ask how many men over 60 had a high-risk waist circumference. From the twenty-five rows it is not, and no amount of arithmetic will recover it.

This is also why a published table is insufficient for reanalysis, a point that returns in Chapter 9 when a paper is read in full.

2 Addresses, ranges, and the dollar sign

Every cell has an address written as a column letter followed by a row number: `C3` is the cell in column C, row 3. The order is fixed and reversing it is the commonest early error.

A cell holds either a value or an instruction. An instruction begins with an equals sign, and

without one the same characters are stored as text and nothing is computed. Everything else rests on the distinction between what a cell displays and what it contains: the cell shows the result, and the formula bar above the sheet shows the instruction that produced it.

A range names a block of cells: `E2:E1811` is one column of a register holding 1,810 participants, `B2:E2` is four variables for one participant, and `B2:E1811` is the whole block. Two errors here produce plausible answers rather than complaints. `E1:E1811` includes the header row, and `E2:E1810` omits the last participant.

Range
Two cell addresses separated by a colon, denoting every cell in the rectangle they bound.

2.1 Relative and absolute references

A reference written as `C2` is *relative*. It is stored not as an address but as a direction: two columns to the left, same row. Copying the formula to the row below therefore produces `C3`, which is what makes it possible to compute an entire column of percentages by writing one formula. It is also what breaks a percentage, because a percentage has a numerator that changes down the column and a denominator that does not.

Prefixing a dollar sign fixes the part that follows it. `C$27` fixes the row, `$C27` fixes the column, and `$C$27` fixes both. On Windows, F4 cycles through the four forms while the formula is open; on macOS the keystroke is fn+F4.

Absolute reference
A reference whose row, column, or both are fixed with a dollar sign, so that copying the formula does not move it.

$$\text{percentage} = \frac{a}{N} \times 100 \tag{2.1}$$

a the count of observations in the category, which changes from row to row
N the denominator, which does not

Equation (2.1) is trivial as arithmetic and is the source of a large share of published error, because the choice of N is a decision about what question is being answered and the software has no way to check it. Written into a spreadsheet, the decision becomes visible: a is relative, N is absolute, and a formula in which N is relative is a formula in which the denominator was not, in fact, decided.

Example 2.1 • A column of percentages, and the same column broken

The Sylhet survey reports 260, 259, 167 and 178 men in four age bands, out of 864 men in total. With the count of men in column C and the total in row 27, the formula `=C2/C$27*100` entered once and copied down returns 30.1, 30.0, 19.3 and 20.6, reproducing the printed column exactly.

Written as `=C2/C27*100` the first row is still 30.1, because the denominator has not yet moved. The second row divides 259 by 178 and returns 145.5, and the third divides by an empty cell and returns `#DIV/0!`. The first row being correct is what makes the error survive: the usual check is to look at the top of a column.

Watch out

A percentage above 100 in a column of proportions is not a rounding problem. It means the denominator moved.

3 Counting a categorical column

Chapter 3 established that describing a categorical variable consists entirely of counting, and that whatever goes wrong goes wrong in the denominator. This section is that chapter performed rather than read.

3.1 Counting under a condition

`COUNTIF` takes two arguments: a range, and a condition that a cell must satisfy to be counted. The condition is written in quotation marks and may use any of the comparisons `>`, `>=`, `<`, `<=` and `<>`. `COUNTIFS` takes the same arguments in pairs, as many pairs as required, and counts the observations satisfying all of them at once. Two pairs produce one cell of a two-way table; four formulas produce the whole of a 2×2.

Example 3.1 • A two-way table, built four cells at a time

In the practice register, sex is column B, coded 1 for male, and smoking is column G, coded 1 for yes. The formula `=COUNTIFS(B2:B1811,1,G2:G1811,1)` returns 546.

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

The margins are computed with `SUM` rather than typed, which makes the table check itself: the row totals must reproduce the two group sizes and the column totals must reproduce the two smoking totals. A table built this way cannot silently disagree with its own margins.

3.2 The same numerator, two denominators

The four counts above support two different percentages, and the difference between them is one cell reference.

Question	Denominator	Answer
Of all smokers, how many are men?	the 582 smokers	93.8%
Of all men, how many smoke?	the 864 men	63.2%

Both are correct arithmetic and only one answers any given question. The first is a row percentage and describes the composition of the smokers; the second is a column percentage and describes the behaviour of the men. A sentence about whether men smoke more than women requires the second. The survey prints the second, as its own footnote states, and Chapter 3 was built on a reader taking it for the first.

Watch out

A third answer is available and is the commonest of all: dividing by the whole sample, 1,810, which gives the share of everybody who is both male and a smoker. It answers a real question, and almost never the one being asked.

3.3 Percentages that do not sum to 100

A column of percentages that sums to 100.1 is usually not an error. Three categories that each round up produce exactly that. The survey's fruit-and-vegetable column reports 49.1, 38.6 and 12.4, summing to 100.1.

What does warrant going back to the counts is a larger discrepancy, or counts that do not sum to their own stated *n*. Chapter 3 recorded one genuine slip of this kind in the same table, where a printed percentage differs from the recomputed one by 0.1: 1,228 non-smokers out of 1,810 is 67.8%, and 67.9% is printed. It changes nothing, and it is the reason that column sums to 100.1 rather than 100.0.

4 Summarising a numerical column

A numerical column is described by its centre, its spread, and the shape implied by the relationship between the two. Chapter 4 treats all three properly. This section produces them.

4.1 Centre and the five-number summary

Five-number summary
The minimum, first quartile, median, third quartile and maximum of a column.

Four functions cover the centre and the extremes, and all four take the same argument: `AVERAGE`, `MEDIAN`, `MIN` and `MAX`. Quartiles come from `QUARTILE.INC`, whose second argument is 1, 2 or 3, and any other percentile from `PERCENTILE.INC`, whose second argument is a proportion. The second quartile and the median are the same quantity under two names.

Example 4.1 • A five-number summary, and what it can be compared with

The practice register's waist column returns a median of 77.0 cm with quartiles 69.7 and 84.8, a mean of 77.3, a minimum of 52.2 and a maximum of 107.9.

The Sylhet survey prints a median waist of 77.0 cm with an interquartile range of 69.7 to 84.8. The three agree because the register was generated to reproduce those printed figures and for no other reason. Agreement checks that the file is what it claims to be. It is not evidence about anybody's waist. The minimum and maximum were chosen when the file was built and correspond to nothing the survey reported.

Watch out

`QUARTILE.INC` and `QUARTILE.EXC` implement two defensible conventions and disagree in the second decimal place. Statistical packages differ again. The difference has

never changed a conclusion, and mixing the two within one table is untidy rather than wrong.

4.2 When the mean and the median separate

Where a column is roughly symmetric the mean and the median nearly coincide. Where it is not, they separate, and the direction of the separation says which way the long tail runs. The register's waist column gives a mean of 77.3 against a median of 77.0. Its schooling column gives a mean of 3.25 against a median of 2, because no one can have fewer than zero years of schooling and a few people have twelve.

Which of the two to report, and what to do when they disagree, is Chapter 4. What is worth carrying from here is only that computing both costs nothing and that the gap between them is information.

4.3 Standard deviation, and which one

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \tag{4.1}$$

- x_i the value for the ith observation
- $\bar{x}$ the mean of the column
- n the number of observations
- s the sample standard deviation, returned by `STDEV.S`
- σ the population standard deviation, returned by `STDEV.P`

Only the divisor differs. `STDEV.S` is the one to use. A study measures a sample so as to describe a population it did not measure, and that is the situation the divisor $n - 1$ is for. Why dividing by n understates the spread, and why $n - 1$ corrects it, is Chapter 7, which demonstrates it by repeated sampling rather than by argument.

Example 4.2 • A choice that looks unimportant and is not

On the full register of 1,810 rows, `STDEV.S` returns 10.02 and `STDEV.P` returns 10.01: the two agree to the second decimal place. On the 40-row extract, with 38 measured values, they return 9.73 and 9.60.

Their ratio is $\sqrt{n/(n-1)}$, which approaches 1 as n grows. So the choice is invisible in a large study and material in a small one, and a small study is where a wrong answer is least likely to be questioned.

4.4 Two quantities built from a mean and a standard deviation

Where a study reports a mean with a standard deviation, two further quantities follow immediately. The coefficient of variation is the standard deviation divided by the mean:

$$\mathrm{CV} = \frac{s}{\bar{x}} \times 100 \tag{4.2}$$

Coefficient of variation The standard deviation expressed as a percentage of the mean, allowing spread to be compared across variables measured in different units.

s the sample standard deviation
$\bar{x}$ the sample mean

Example 4.3 • Coefficient of variation in a cluster trial

A cluster-randomised trial of hypertension management in rural South Asia reports a baseline systolic blood pressure of 146.7 ± 22.4 mm Hg in its intervention group and a mean age of 58.8 ± 11.5 years. The coefficients of variation, computed here and not printed by the trial, are 15.3% and 19.6%.

Note the reversal. Systolic pressure has much the larger standard deviation and much the smaller coefficient of variation, because it is measured on a scale whose values are large. One SD either side of the systolic mean spans 124.3 to 169.1 mm Hg, which is also our arithmetic.

4.5 Weighted sums

`SUMPRODUCT` multiplies two ranges element by element and adds the results, which is what a weighted average requires:

$$\bar{p} = \frac{\sum_{k} w_k \, p_k}{\sum_{k} w_k} \tag{4.3}$$

w_k the weight given to the kth stratum, here the share of a standard population falling in an age band
p_k the rate observed in that stratum
$\bar{p}$ the standardised rate

Example 4.4 • Direct standardisation in one formula

Chapter 3 compared hypertension prevalence between two waves of a national survey and asked whether the rise could be an artefact of an ageing population. The survey prints prevalence in five age bands together with the age structure of the two waves combined. Standardising each wave to that common structure is one application of equation (4.3), which in a spreadsheet is `=SUMPRODUCT(B2:B6,C2:C6)/SUM(B2:B6)`.

Crude, the two waves give 25.84% and 39.40%. Standardised, computed here, they give 25.91% and 39.29%, a ratio of 1.52. The rise survives the adjustment essentially untouched, so age structure does not explain it. A check that returns no difference is a result, not a wasted step.

5 The function reference

Every function this chapter uses, with the syntax Excel expects. A range written `E2:E1811` assumes a register of 1,810 participants with a header on row 1.

Function	Worked cell	What it returns
`COUNTA`	`=COUNTA(A2:A1811)`	cells that are not empty
`COUNT`	`=COUNT(E2:E1811)`	cells holding a number
`COUNTIF`	`=COUNTIF(C2:C1811,">=65")`	cells meeting one condition
`COUNTIFS`	`=COUNTIFS(B2:B1811,1,G2:G1811,1)`	cells meeting every condition given
`SUM`	`=SUM(H21:H23)`	the total of a range
`AVERAGE`	`=AVERAGE(E2:E1811)`	the arithmetic mean, ignoring blanks
`MEDIAN`	`=MEDIAN(E2:E1811)`	the middle value
`MIN`, `MAX`	`=MIN(E2:E1811)`	the extremes
`QUARTILE.INC`	`=QUARTILE.INC(E2:E1811,1)`	the first quartile; 2 gives the median
`PERCENTILE.INC`	`=PERCENTILE.INC(E2:E1811,0.9)`	the 90th percentile
`STDEV.S`	`=STDEV.S(E2:E1811)`	the sample standard deviation, divisor $n-1$
`STDEV.P`	`=STDEV.P(E2:E1811)`	divisor n; rarely the right choice
`SUMPRODUCT`	`=SUMPRODUCT(B2:B6,C2:C6)`	the sum of element-by-element products

A reference to another sheet is written with the sheet name, an exclamation mark, and then the address: `=MEDIAN(cohort!E2:E1811)`. A summary kept on its own sheet, pointing at the data sheet, is a tool that can be reused rather than a calculation to be repeated.

6 Error values, and what each one means

An error value is a better outcome than a wrong number, because it can be seen.

Shown	Cause
`#DIV/0!`	The denominator is zero or empty. In a percentage column this usually means a relative reference has been copied past the row holding the total.
`#VALUE!`	Arithmetic has been attempted on text. A cell reading `88 cm` or `52 yrs` is text, and so is a number typed with a stray space.
`#NAME?`	The function name is misspelt, or a text condition has been written without quotation marks.
`#REF!`	The formula points at a cell that no longer exists, usually because a row or column was deleted after the formula was written.
`#####`	Not an error. The column is too narrow to display the number.

Two failures produce no error value at all and are correspondingly more dangerous. A number stored as text is simply excluded from every calculation, so a mean is computed over fewer observations than the analyst believes. A sentinel value such as 999 is included in every calculation, so a mean is computed over a value that was never measured.

7 Provenance of the figures in this chapter

Three classes of number appear above and are distinguished throughout.

Reported figures were printed by a study: the counts and percentages of the Sylhet survey's Table 1 and Table 2, the baseline figures of the cluster trial, and the age-band prevalences of the national survey.

Ours are computed here from what those studies printed: 339 hypertensive participants as the sum of the two sex-specific counts; 93.8% as 546 divided by 582; 67.8% as 1,228 divided by 1,810; the coefficients of variation 15.3% and 19.6%; the band 124.3 to 169.1; and the standardised prevalences 25.91% and 39.29% with their ratio 1.52.

Simulated figures come from the practice register, which was generated rather than collected, because no study in this course publishes participant-level data. Its waist, age and schooling columns were matched to the printed quartiles of the Sylhet survey, and its sex, smoking, BMI band and hypertension columns to the printed counts within each sex. Its tails, and every relationship between columns not in that list, were chosen when the file was built. No figure from it is evidence about any population.

8 Chapter summary

1. A dataset holds one row per unit of observation and one column per variable, with variable names on row 1 and nothing but data below.
2. Categorical variables are recorded as codes, with the meanings kept in a codebook in

the same file. A code does not create an order where none exists.

3. Missing values are left blank. A blank is excluded from every summary function; a sentinel number is included in all of them.
4. `COUNTA` minus `COUNT`, computed for each variable, is the missing-data count STROBE asks a paper to report.
5. A cell reference is stored as a direction, not an address, which is why copying a formula moves it. A dollar sign fixes the part that follows.
6. A percentage has a numerator that varies and a denominator that does not, so the denominator is written absolute. A column of percentages containing a value above 100 means the denominator moved.
7. `COUNTIF` counts under one condition and `COUNTIFS` under several. Four `COUNTIFS` formulas and two `SUM` formulas produce a 2×2 table that checks its own margins.
8. The same two counts give a row percentage or a column percentage depending on the denominator chosen, and only the question decides which is wanted.
9. A column of percentages summing to 100.1 is usually three honest round-ups.
10. `AVERAGE`, `MEDIAN`, `MIN`, `MAX` and `QUARTILE.INC` give the centre and the five-number summary, all from the same range.
11. `STDEV.S`, with divisor $n-1$, is the standard deviation a study reports. The alternative differs negligibly in a large sample and materially in a small one.
12. `SUMPRODUCT` divided by `SUM` is a weighted average, which is what direct standardisation requires.

9 Key terms

Absolute reference A cell reference whose row, column or both are fixed with a dollar sign, so that copying the formula does not move it.

Codebook The record of what each code in a dataset means, kept with the data.

Coefficient of variation The standard deviation as a percentage of the mean, written CV.

Column percentage A percentage whose denominator is the total of its own column in a cross-tabulation.

Five-number summary Minimum, first quartile, median, third quartile and maximum.

Range Two cell addresses separated by a colon, denoting the rectangle they bound.

Relative reference A cell reference stored as a direction from the formula's own position, and therefore moved when the formula is copied.

Row percentage A percentage whose denominator is the total of its own row in a cross-tabulation.

Sentinel value A number such as 999 used to record a missing observation. It is indistinguishable from a measurement.

Unit of observation The entity one row describes.

s The sample standard deviation, divisor $n - 1$.

σ The population standard deviation, divisor n.

10 Exercises

Exercises 9 to 14 use the workbooks distributed with this chapter. Those marked with a study use figures printed in the papers listed in section 12.

Part A. The shape of a dataset

1. A ward register records, for each of 40 patients, the admission date, three blood pressure readings taken on different days, and the discharge outcome. State how many rows and how many columns a properly shaped dataset of these records has, and say what the unit of observation is.

2. A sheet records smoking status as *Yes*, *yes*, *YES*, *Y*, *smoker* and *Yes, daily*. Give the count that `=COUNTIF(D2:D101,"Yes")` returns if these six spellings are equally common among 60 smokers in 100 patients, and say what the correct recording would have been.

3. Explain why a missing waist measurement recorded as 999 is more dangerous than one recorded as a blank cell, and name the one summary statistic most likely to reveal it.

Part B. References and denominators

4. A formula in cell `F2` reads `=C2/C27*100`. State what it reads after being copied to `F5`, and what it should have read.

5. The first row of a percentage column is correct and the rest are not. Explain why this pattern is characteristic of a missing dollar sign rather than of a mistyped denominator.

6. A colleague computes a percentage column with `=C2/864*100`, typing the denominator directly. The formula is correct today. Give one circumstance in which it silently becomes wrong.

Part C. Reading a result

7. In a sample of 1,810 adults, 582 smoke and 546 of those are men, out of 864 men in the sample. Compute both percentages that these figures support, and state which of the two belongs in a sentence about whether men smoke more than women.

8. A column of three percentages reads 49.1, 38.6 and 12.4. Give the sum, say whether it indicates an error, and state what would.

Part D. Calculation

9. Using the `extract` sheet of the practice register, give the values returned by `=COUNTA(A2:A41)` and `=COUNT(E2:E41)`, and explain the difference between them.

10. Using the `cohort` sheet, write the formula that returns the number of participants aged 65 or over, and the formula that returns that number as a percentage of everyone with an age recorded.

11. Write the five formulas of a five-number summary for the waist column of the `cohort` sheet, placed on a separate sheet.

12. A trial reports a baseline systolic blood pressure of 146.7 ± 22.4 mm Hg. Compute the coefficient of variation and the interval one standard deviation either side of the mean, and state which of the three figures is the trial's and which are yours.

13. A national survey reports hypertension prevalence in five age bands, together with the share of the population in each band. Write the single formula that returns the age-standardised prevalence, and name the quantity it computes.

Part E. Appraisal

14. A published table reports a mean waist circumference of 131 cm in a sample of rural adults, with a maximum of 999. State the most likely explanation and what should have been reported instead.

15. A reader reproduces a paper's published median exactly from a dataset generated to match that paper's printed quartiles, and concludes that the paper's result is confirmed. Explain what the reproduction does establish and what it does not.

11 Solutions

1. Forty rows, one per patient, which is the unit of observation. The columns are the admission date, three separate columns for the three readings, and the discharge outcome: six columns, plus a patient identifier, giving seven. The three readings occupy three columns rather than three rows because they describe the same patient.

2. Ten. `COUNTIF` matches text exactly but ignores capitalisation, so *Yes* and *yes* and *YES* all match, giving thirty of the sixty; *Y*, *smoker* and *Yes, daily* do not match at all. The count returned is therefore 30, not 60. The correct recording is a single code, `1` for a smoker and `0` for a non-smoker, with the meaning recorded in a codebook.

3. A blank cell is excluded from `COUNT`, `AVERAGE`, `MIN`, `MAX` and every other summary function, so it changes the denominator and nothing else. A 999 is included in all of them, so it inflates the mean, the maximum and the standard deviation while leaving the count apparently complete. The maximum is the statistic most likely to reveal it, because a waist circumference of 999 cm is visibly impossible, whereas an inflated mean is merely surprising.

4. After copying to `F5` it reads `=C5/C30*100`: both references moved by three rows. It should have read `=C5/C$27*100`, with the denominator fixed by the dollar sign so that only the numerator moves.

5. Because a relative denominator is correct in the row where the formula was first written and wrong in every row after it. A mistyped denominator is wrong everywhere, including the first row, and is therefore visible immediately. The characteristic signature of the missing dollar sign is a correct first row followed by values that drift and then fail, ending in `#DIV/0!` once the reference passes the last populated cell.

6. As soon as a row is added to or deleted from the data. The typed 864 does not change, so every percentage in the column is computed against a denominator that no longer describes the sample. Writing the denominator as `COUNT(B2:B1811)` or as a reference to a cell holding a `SUM` avoids this.

7. 546 divided by 582 is 93.8%, the share of smokers who are men. 546 divided by 864 is 63.2%, the share of men who smoke. The second belongs in a sentence about whether men smoke more than women, because that sentence compares behaviour between the sexes and therefore needs each sex as its own denominator. The first describes the composition of the smokers and would be the same even if there were very few men in the sample.

8. The sum is 100.1. It does not indicate an error: three percentages that each round up produce exactly this. What would indicate an error is a discrepancy of more than a point or two, or counts that fail to sum to the stated *n* of the column.

9. `=COUNTA(A2:A41)` returns 40 and `=COUNT(E2:E41)` returns 38. The first counts every non-empty cell in the identifier column, and every patient has an identifier. The second counts numeric cells in the waist column, and two patients have no waist measurement. The difference is the missing-data count for that variable.

10. The count, and then the same count as a percentage:

```
=COUNTIF(C2:C1811,">=65")
=COUNTIF(C2:C1811,">=65")/COUNT(C2:C1811)*100
```

They return 338 and 18.7. The denominator is written as `COUNT` rather than as 1,810 so that it follows the data if rows are added or removed.

11. With the data on a sheet named `cohort`:

```
=MIN(cohort!E2:E1811)
=QUARTILE.INC(cohort!E2:E1811,1)
=MEDIAN(cohort!E2:E1811)
=QUARTILE.INC(cohort!E2:E1811,3)
=MAX(cohort!E2:E1811)
```

They return 52.2, 69.7, 77.0, 84.8 and 107.9. The median could equally be written `=QUARTILE.INC(cohort!E2:E1811,2)`.

12. The coefficient of variation is $22.4 / 146.7 \times 100 = 15.3\%$. One standard deviation either side of the mean spans 124.3 to 169.1 mm Hg. The mean and the standard deviation are the trial's, printed in its baseline table; both derived figures are yours, and should be labelled as such wherever they appear.

13. `=SUMPRODUCT(B2:B6,C2:C6)/SUM(B2:B6)`, where column B holds the share of the standard population in each band and column C the prevalence observed in that band. It computes the age-standardised prevalence: the prevalence the population would show if its age structure were that of the standard population. Applied to the two waves of the national survey it returns 25.91% and 39.29%.

14. A sentinel value of 999, used to record a missing measurement, has been included in the mean. The reported maximum gives it away. What should have been reported is the mean computed over the measured values only, with the number of missing observations stated separately, as STROBE item 14(b) requires.

15. It establishes that the generated dataset is what it claims to be: the generation procedure did reproduce the printed quartiles. It establishes nothing whatever about the population the paper studied, because the dataset contains no information from that population beyond the three summary figures it was built to match. A number computed from such a file is a demonstration of a calculation, never a finding.

12 References and further reading

Every entry carries a PMID or a DOI. Where none exists, the entry says so.

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.
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. Chowdhury MAB, Islam M, Rahman J, Uddin MT, Haque MR, Uddin MJ. Changes in prevalence and risk factors of hypertension among adults in Bangladesh: an analysis of two waves of nationally representative surveys. *PLoS One* 2021;16(12):e0259507. PMID 34855768. PMC8638884. doi:10.1371/journal.pone.0259507.
4. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement. *J Clin Epidemiol* 2008;61(4):344–349. PMID 18313558.
5. Daniel WW, Cross CL. *Biostatistics: A Foundation for Analysis in the Health Sciences*, 9th edition. Wiley. A textbook; no PMID or DOI exists for it.

13 For students in the mentorship programme

Before the next session

1. Rebuild the five-number summary of exercise 11 from scratch, on a new sheet, without consulting the solution.
2. Repair the `messy` sheet of the practice register into something that can be computed on. Write down what was decided about the smoking column and why.
3. Compute both percentages of exercise 7 in the workbook, and write one sentence for each saying what question it answers.

Completed answers are in `S3-5-05-completed.xlsx`. It is worth attempting each task before opening it.

Reading

Daniel, chapter 2, sections 2.1 to 2.5, book pages 19 to 48. These cover the statistics behind what this chapter computed, and are the reading for the next session rather than for this one.

Questions this chapter leaves open

Question	Where it is settled
When is a mean the right summary and when is a median?	Chapter 4, describing numerical data
What does a standard deviation actually describe?	Chapter 4
Why is the divisor $n - 1$ rather than n?	Chapter 7, by repeated sampling
How is any of this drawn rather than tabulated?	Chapter 6, tables and graphs
What shape does a column have, and when is a value an outlier?	Chapter 5, distributions and outliers

The next session takes the five-number summary produced here and asks what each of its parts is for.