Stat.BD

APPLIED BIOSTATISTICS & RESEARCH METHODOLOGY • SESSION 2

Population, Samples, Variables & Data

What one row is, what one column is, and what job each column does

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

Session 2 of 30

Where we left it

Last session ended on one distinction.

- **Description** is a statement about the people you measured.
- **Inference** is a statement about the people you did not.

Everything today sits underneath that sentence. Inference needs two things named before it can start: *who* you are talking about, and *what* you measured on them.

The gap this session fills

You can now tell description from inference. You cannot yet say precisely *what* is being inferred, *about whom*, or *from which column*. Those three gaps close today.

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

The test of this session

Take the paper you brought. Write out every variable in it, one per row, saying for each: what it is called, what job it does in the study, what kind of thing it is, how it was measured, and in what unit.

That table is not an exercise. It is the first page of your own protocol, and you will write one for every study you ever design.

- Who the study is about: **population** and **sample**
- What was measured: **variables**, and what kind each one is
- What job each measurement does: **roles**
- Why the job depends on the question: **association and causation**

Act I

Your question, turned into rows and columns

Stat.BD

Your question

You brought a question from your own practice.

It goes on the board now, and it stays there for ninety minutes.

Four slots to fill

- P who the question is about
- I what is done to them
- C compared with what
- O measured how

The standard, from a paper you already know

P adults with hypertension in rural South Asia
I home visits by community health workers
C usual care
O systolic blood pressure at 24 months

COBRA-BPS. Jafar TH et al. *N Engl J Med* 2020;382(8):717–726. PMID 32074419.

Every study, in the end, becomes a rectangle

one column is one variable
the same thing measured on everybody

one row is one participant
the unit of observation

ID	Age	Sex	Waist (cm)	SBP	Hypertension
001	52	M	88.0	138	No
002	67	F	84.5	152	Yes
003	41	M	72.3	118	No
004	59	F	91.2	146	Yes

this whole object is "the rectangle"

Illustrative rows, in the shape of the Sylhet survey dataset. Whatever the design, this is what reaches the analysis. Before anything else can be said about a study, two questions have answers: *what is one row?* and *what is one column?*

Two questions to ask before any other

1. What is one row?

A patient? An admission? A pregnancy? A tooth? A village?

The answer decides what *n* counts, and it is not always a person.

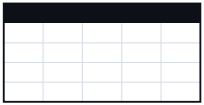
2. What is one column?

One thing, measured the same way, on everybody in the study.

If it was measured differently on different people, it is not one column.

Where this goes wrong in practice

A study of 200 eyes in 120 patients does not have 200 independent patients. A study of 500 visits by 90 children does not have 500 children. The rectangle tells you; the abstract often does not.



Both questions are questions about *this* object, the one from four slides ago. Everything today is a claim about one of its rows or one of its columns.

So what is one row?

Sylhet survey

Wealth is a property of the *household*. Blood pressure is a property of the *person*. One row is a person.

COBRA-BPS trial

30 villages were randomised. 2,645 people were measured. One row is a person; the treatment was not.

SDG panel

114 countries, each observed in 25 years. One row is a *country-year*, and nobody in it is a person.

Say what one row is before you say anything else about a study. The answer decides what the sample size means, and it is not always a patient.

The row you measured is not always the row you treated

In COBRA-BPS, thirty villages were randomised and 2,645 people were measured.

- The **treatment** was given to a village.
- The **blood pressure** belongs to a person.

So one row is a person, and the thing that was assigned sits a level above the row.

Why it is not a technicality

Two people in the same village got the same treatment, from the same worker, on the same day. Their readings are not two independent pieces of evidence.

Counting them as though they were makes the study look more certain than it is.

Jafar TH et al. *N Engl J Med* 2020;382(8):717–726. PMID 32074419.

And what is one column?

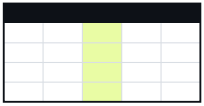
One thing, recorded *the same way*, on everybody in the study.

- Not ”blood pressure”.
  Systolic, in mm Hg, the average of the second and third readings.
- Not ”smoking”.
  Currently smoking, yes or no, as reported by the participant.

When it is not one column

If half the participants were measured with a calibrated machine and half were asked to remember, that is not one column with some noise in it.

It is two measurements stacked into one place, and the stacking is a decision somebody should have defended.



One column. Not one reading, and not one kind of reading: one procedure, applied the same way to everybody in the rectangle.

Act II

Who the study is actually about

Stat.BD

Why not simply measure everybody?

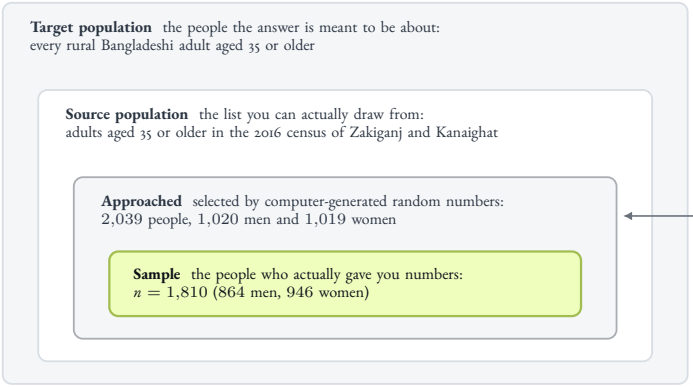
The Sylhet team already had a list of every adult in two subdistricts. They still measured only 2,039 of them.

- **Cost and time.** Three readings, a tape and a questionnaire, per person, at home.
- **Quality.** A small survey done carefully beats a large one done badly.
- **Measuring sometimes destroys.** Nobody tests every parachute.
- **Some of them do not exist yet.** The answer has to serve next year's patients too.

What sampling actually buys

Not a shortcut. A study you can finish, done well enough to believe.

Four populations, not one



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

Who was let in, and who was kept out

Inclusion criteria

What a person must be to enter the study.

Here: aged 35 or older, living in one of the two subdistricts, on the 2016 census list.

Exclusion criteria

What disqualifies somebody who otherwise qualifies.

Here: pregnancy. 14 women were excluded on that ground.

They are not opposites

An exclusion is not just the reverse of an inclusion. It is a separate decision, made for a separate reason, and it needs a separate justification.

Pregnancy changes blood pressure. Excluding pregnant women makes the remaining measurements cleaner and makes the answer not about them.

Every criterion narrows the population the answer applies to. That is the cost, and it is paid whether or not anybody notices.

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

How those 2,039 were picked

1. A census of both subdistricts already existed, listing every adult aged 35 or older.
2. Names were drawn from it by **computer-generated random numbers**.
3. Fieldworkers went to those households, and to no others.

What the randomness is for

Not fairness. It is the one procedure that gives no one a reason to be in the sample that is also a reason to have high blood pressure.

The alternative, and why it fails

If the fieldworker chooses, he chooses the house on the road, the family he knows, the man who is at home in the afternoon. Every one of those is a reason to be in the sample that is also a reason to have a particular blood pressure.

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

Who was never on the list at all

The sampling frame was a census of the two subdistricts taken in 2016. Fieldwork ran from August 2017.

So anybody who moved into those subdistricts after the census was never eligible to be drawn, however long they had lived there by the time the fieldworkers arrived.

They are not non-responders. They are invisible: they never appear in any figure, including the ones that count who was lost.

Why this one is hard to see

Non-response is countable, and this paper counts it honestly.

A gap in the sampling frame is not countable from inside the study, because the study's only picture of the population is the frame itself.

Census year and fieldwork dates from Khanam 2019, Methods. PMID 31662350. The inference about recent arrivals is ours.

The 229 people who never arrived

Of 2,039 people approached, 1,810 completed.

Refused	36
Absent	56
Migrated out	76
Died	42
Pregnant (excluded)	14
Incomplete data	5

Why this is not bookkeeping

Ask of each line: could the reason a person is missing be related to the thing being measured?

42 people died before the survey reached them. Blood pressure is not unrelated to dying.

Counts summed by us from the losses reported for men and women separately. Khanam 2019, PMID 31662350.

A parameter

A number that describes the **population**.

Fixed. Nobody has ever seen it, and nobody ever will.

π, the proportion of rural Bangladeshi adults with hypertension.

A statistic

A number that describes the **sample**.

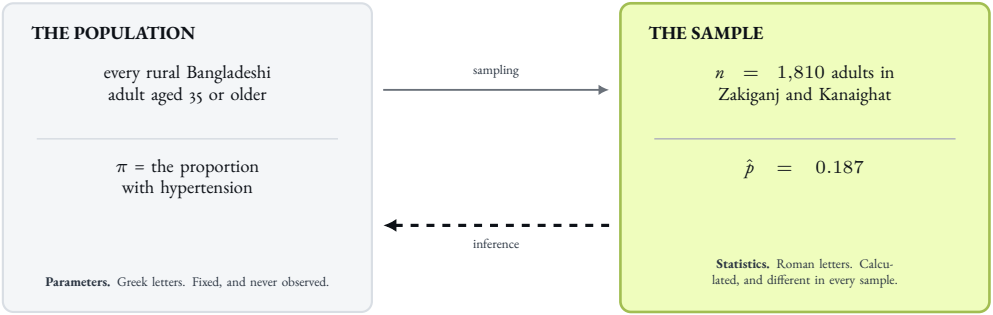
Calculated. Different in every sample you could have drawn.

$\hat{p} = 0.187$, from the 1,810 adults who were measured.

The reason the pair matters

Every result in every paper you will ever read is a statistic. Every question worth asking is about a parameter. The rest of this course is the distance between them.

Two worlds, and one risky arrow between them



Everything measured lives on the right. Everything the study is actually about lives on the left. The dashed arrow is the whole of statistical inference, and it is the only one of the two that can be wrong.

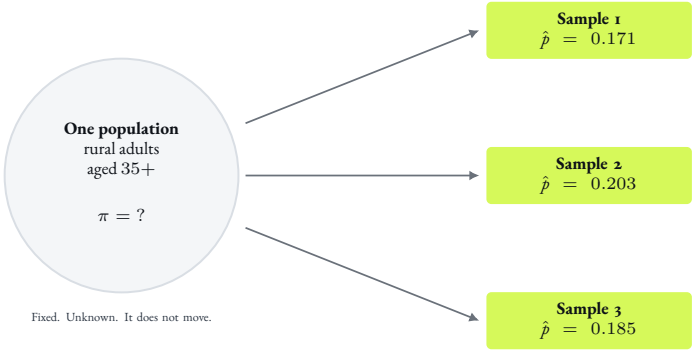
The notation, once, so it stops being decoration

Population		Sample	
N	size	n	size
μ	mean	$\bar{x}$	mean
σ	standard deviation	s	standard deviation
π	proportion	$\hat{p}$	proportion
Greek. Fixed. Never seen.		Roman. Calculated. Different every time.	

The whole convention in one line

If it is Greek, nobody has ever seen it. If it is roman, somebody computed it from data.

Three samples, three answers



Three samples, three answers.

Nothing about the population changed between them. Only the people who happened to be drawn changed.

Illustrative values, to show that a statistic moves and a parameter does not. How far it moves, and how to say so, is a later session.

One substitution, so the symbols have somewhere to land

Of 1,810 adults surveyed, 339 had hypertension.

$$\hat{p} = \frac{339}{1{,}810} = 0.187$$

The paper reports it as 18.7%, with a 95% confidence interval of 17.0 to 20.6.

What each piece is

339 and 1,810 are counts, from the study.

0.187 is $\hat{p}$, a statistic.

π, the proportion in rural Bangladesh, is still unknown and always will be.

Counts and percentage from Khanam 2019, PMID 31662350. The division shown here is ours.

Act III

What kind of thing is this column, and what job does it do?

Back to the rectangle, because the rest of today lives in it

1. Every column has a *kind*
binary, nominal, ordinal, discrete or continuous

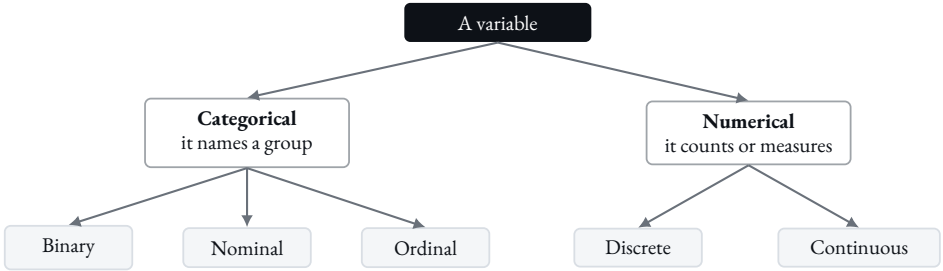
2. Every column has a *job*
outcome, exposure, predictor or covariate

ID	Age	Sex	Waist (cm)	SBP	Hypertension
001	52	M	88.0	138	No
002	67	F	84.5	152	Yes
003	41	M	72.3	118	No
004	59	F	91.2	146	Yes

3. Every cell is one value
produced by one procedure, on one participant

The kind is a fact about the column. The job is a fact about the question. **Only one of the two moves**, and the rest of today is about which one.

What kind of thing is this column?



The boundary that is genuinely arguable: ordinal or discrete?

A pain score out of ten.

The case for ordinal

The gap between 2 and 3 is not the same amount of pain as the gap between 8 and 9. Nobody calibrated the scale, so averaging it averages unequal things.

The case for discrete

It is a whole number on a fixed range, everybody uses the same anchors, and treating it as numerical lets you use methods that are far more informative.

What actually decides it

Not the column. The *question*, and the amount of trouble you are willing to defend.

Reporting a median pain score of 6 is unarguable. Reporting a mean of 6.3 is a claim that the scale has equal intervals, and somebody may ask you to defend it.

The tree from two slides ago is a tool for choosing a summary, not a taxonomy anybody defends to the death.

Sort these eight

1. Age, in completed years
2. Sex
3. Years of schooling
4. Wealth tertile
5. Waist circumference, in cm
6. Current smoker
7. Servings of fruit and vegetables per day
8. Hypertension

Binary • Nominal • Ordinal • Discrete • Continuous

All eight are columns of the Sylhet survey. Khanam 2019, PMID 31662350.

The answers, and the three that are worth arguing about

Column	Kind	Why it is worth a second look
Sex	Binary	Two groups, no order.
Current smoker	Binary	Yes or no, as reported by the participant.
Wealth tertile	Ordinal	Ordered, unequal gaps.
Hypertension	Binary	Not measured. Built from two other columns.
Waist circumference	Continuous	Measured in cm, to one decimal place.
Age	Discrete	Counted in whole years, treated as continuous.
Years of schooling	Discrete	Counted, and often used as if ordered bands.
Fruit and vegetables	Discrete	Counted servings, reported by the participant.

The same variable, twice, in two different types

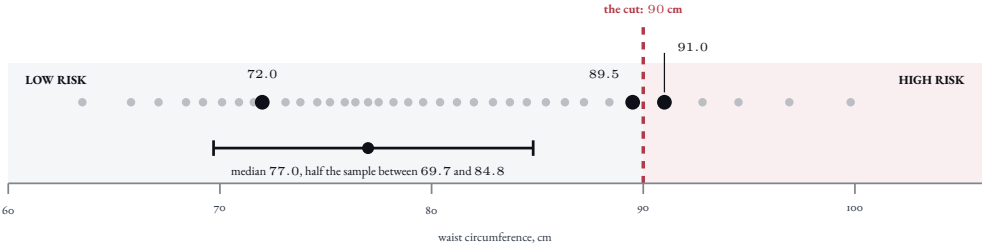
	As measured	**As reported**
Body mass index	20.3 kg/m^2, median	Under / normal / overweight
Waist circumference	77.0 cm, median	Low risk / high risk
Years of schooling	2 years, median	None / 1 to 5 / 6 or more
Age	48 years, median	35–44 / 45–54 / 55–64 / 65+

What is going on

These are not four variables and their four summaries. They are four variables, each appearing twice in the same table: once as the number that was recorded, and once as a set of boxes somebody decided to sort it into.

Medians for the whole sample. Khanam 2019, Table 1. PMID 31662350.

What a threshold costs you



72.0 and 89.5: the same.
17.5 cm apart, and the categorised column cannot tell them apart at all.

89.5 and 91.0: different.
1.5 cm apart, which is inside the variation between two careful tape measurements.

The reported odds ratio of 4.0 is for the two boxes, not for the centimetres. The cut buys a rule a health worker can apply with a tape measure and no calculator, and it pays for that with every distinction inside each box.

Khanam 2019, Tables 1 and 3. PMID 31662350. Dot positions illustrative; median, IQR, cut and odds ratio as reported.

What the cut bought

What was given up

Every distinction inside each box, and a false one at the boundary.

What was bought

A rule a health worker can apply with a tape measure and no calculator, in a village, today.

So it is a trade, not a mistake

The authors chose waist circumference over body mass index partly because it is easier to measure in the field. A study meant to inform a screening programme has good reason to report the quantity the programme will actually use.

What makes it a problem

Silence. A reader who meets an odds ratio of 4.0 and assumes it refers to centimetres has been misled, whether or not anybody intended it.

Where a column comes from

MEASURED

an instrument

↓

Waist circumference
a tape, in cm

REPORTED

the participant

↓

Current smoker
what he said

CONSTRUCTED

a rule

↓

Hypertension
built from other columns

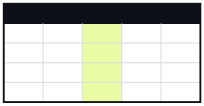
↓

systolic $\geq$ 140 **or** diastolic $\geq$ 90
or taking antihypertensive medication

For every column in any paper, ask which of the three it is.

A constructed column carries every decision that went into constructing it, and those decisions live in the Methods section.

Nobody was measured for this column. Move 140 to 130 and the prevalence changes without one participant changing.



Measured, reported or constructed. Whichever it was, it arrives in the rectangle looking exactly the same, and only the Methods section can tell you which.

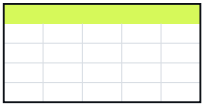
Khanam 2019, Methods. PMID 31662350.

Four jobs a column can do

Outcome	The thing you are trying to explain or predict. What the study is about.
Exposure	The thing whose effect you are asking about.
Predictor	Something used to forecast the outcome, with no claim that it causes anything.
Covariate	Something measured and accounted for so that the exposure can be seen more clearly.

In the Sylhet survey

Outcome: hypertension. Exposure: waist circumference. Covariates: age, wealth, smoking, physical activity.



A job is attached to the column heading, not to the values underneath it. Nothing in the cells tells you which job a column has.

Exposure is a causal word. Predictor is not.

PREDICTION

What is likely to happen to this patient?

A useful flag. Scrubbing the fingers changes nothing at all.

CAUSATION

If I change this, does the outcome change?

Stopping changes the risk. This is what justifies telling a patient to stop.

Both columns predict the outcome, and only one of them is worth intervening on. Calling a column the *exposure* says you believe it is the second kind.

Why a clinician should care

Prediction tells you whom to watch. Causation tells you what to do. Only one of them justifies a prescription.

Act IV

The role is not in the column. It is in the question.

The same columns, two questions

Does a large waist raise the risk of high blood pressure?	
Waist circumference	exposure
Physical activity	covariate
Hypertension	outcome
Age	covariate
Sex	covariate

Does physical activity lower the risk of high blood pressure?	
Waist circumference	covariate
Physical activity	exposure
Hypertension	outcome
Age	covariate
Sex	covariate

Not one number in the dataset changed. The question changed. Waist circumference is the exposure on the left and a covariate on the right. A role is something a question assigns to a column, not a property the column carries around with it.

Both models are the paper's own. Khanam 2019, Table 3 and Results. PMID 31662350.

Association, and what it is not

Association

Two columns move together. Knowing one tells you something about the other.

This is a statement about a dataset, and it can be checked.

Causation

Intervening on one would change the other.

This is a statement about the world, and no dataset contains it.

On the word correlation

Correlation is one narrow way of measuring one narrow kind of association: two numerical columns, moving in a straight line.

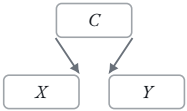
Every correlation is an association. Most associations are not correlations.

Four reasons two columns move together

1. Chance

Nothing links them. This sample simply came out that way, and the next one will not.

2. A common cause



Something else moves both. *X* and *Y* travel together without either touching the other.

3. Reverse causation

The arrow runs the other way. Having the outcome changed the exposure.

4. A real effect

Change *X* and *Y* changes with it. This is the only one of the four that is causation.

An association is the finding. These four are the explanations, and the data alone cannot tell you which one you are looking at. The design is what narrows the list.

In this study, smoking looks protective

Among men in the Sylhet survey, current smokers had *lower* odds of hypertension.

	Odds ratio	**95% CI**
Unadjusted	0.5	0.4 to 0.7
Adjusted	0.7	0.5 to 1.0

Nobody believes smoking prevents hypertension.

So which of the four is it?

The authors raise reverse causation themselves: people who already knew they had high blood pressure may have changed their habits.

Then they argue against it, because about half the people found to have hypertension did not know they had it.

32.2% of this sample smoked, and among men 63.2%. Whatever is going on in this table, it is not a rare exposure.

Khanam 2019, Table 3 and Discussion. PMID 31662350.

Close

Now write it down.

This is what you are going to write

Name	Role	Kind	How a value gets into it	Unit or levels
Hypertension	outcome	binary	$\geq$ 140/90, or on medication	yes / no
Waist circumference	exposure	continuous	tape, standing, to 0.1 cm	cm
Age	covariate	discrete	reported, completed years	years
Wealth tertile	covariate	ordinal	5 household items, scored, cut in thirds	tertile
Physical activity	covariate	ordinal	STEPS questionnaire, MET-min per week, *then cut* into 3 bands	3 bands

Five columns, one row per variable. This is the first page of a protocol, and it is what the session is for.

Kind is a fact about the column. Role is a fact about the question. Only Role moves.

Fill Role in twice, for two different questions, and whatever changes between them is the whole of today.

Physical activity is the row worth a second look: it was recorded as a number and reported as bands, and both belong in the table.

Rows from Khanam 2019. PMID 31662350. Roles are those of the study's main analysis.

The variable table has a professional name

What you are about to build is a **data dictionary**, sometimes called a codebook. Every serious dataset has one, and the ones that do not are the ones nobody can reuse.

Beyond the five columns already named, a working data dictionary adds:

- the variable name as it appears in the file, spelled exactly
- the permitted values, and the code for each
- the code used for missing

The one that causes real damage

A missing value coded as 9, or 99, or -1, in a column where those are also possible real values.

Nobody notices until a mean parity of 9.4 appears in a table.

Write it before you collect anything. Written afterwards it is documentation; written first it is a protocol.

Build the variable table for your own study.

One row per variable. Five columns:

1. Name
2. Role: outcome, exposure, predictor or covariate
3. Kind: binary, nominal, ordinal, discrete or continuous
4. How a value gets into it: the procedure, or the rule that constructs it
5. Unit, or the levels if it has no unit

Next: Session 3, Describing Categorical Data.

Then do it once more

Write down a second question you could ask of the same data, and fill in the Role column again.

Whatever changes between the two is the whole of today.

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, 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. Khair Z, Rahman MM, Kazawa K, et al. Health education improves referral compliance of persons with probable diabetic retinopathy: a randomized controlled trial. *PLoS ONE* 2020;15(11):e0242047. PMID 33180863. PMC7660573. doi:10.1371/journal.pone.0242047.
4. Fahim MMM, Imran MJH, Molla MN, Debnath L, Shil T, Pranto EB, Likhon MMR, Saad MSS, Karim MR. Drivers, receivers, and dynamic linkages: the directed structure of SDG interdependence, 2000–2024. arXiv:2601.20875 [stat.AP], 2026. **Preprint; under review at *Scientific Reports*.**
5. Daniel WW. *Biostatistics: A Foundation for Analysis in the Health Sciences*, 9th ed. Wiley. Sections 1.2 to 1.4, pp. 3–13.