5.1 Data and sampling

Name: _____ Class: _____ Date: _____ **Total: 16 marks**

KEY WORDS sample • bias • data dictionary

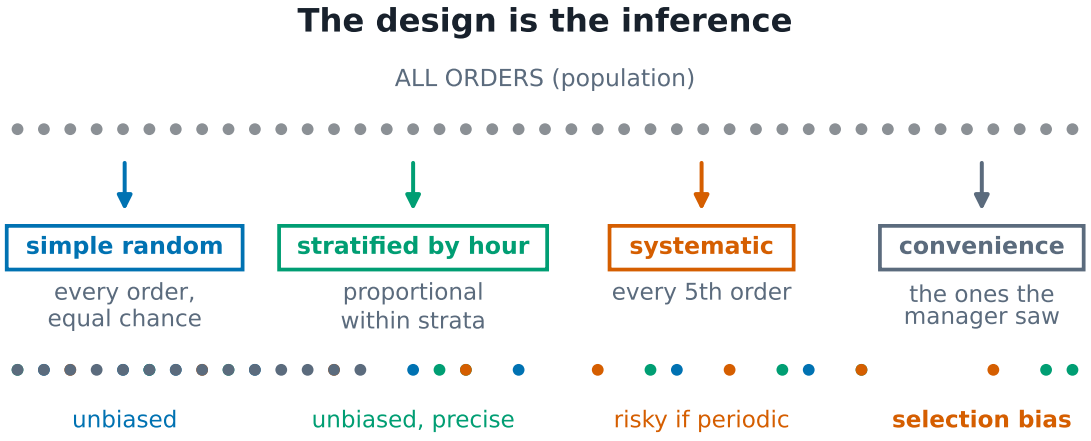
Objective

Classify data types and scales; choose sampling designs to match claims; read a data dictionary and its missingness critically.

1 Key knowledge and worked examples

The design is the inference

A **sample** supports claims about a **population** only through its design: random sampling earns population claims; **stratified** sampling earns comparisons. **Bias** is systematic and survives any sample size. The **data dictionary** records units, codings and what missing means.



size cannot cure bias: 10,000 convenience orders still say nothing about lunch

What you sampled is what you can claim about

■ **Worked example. Nimbus's weekend question**

"Weekends slower?" needs 20 Saturday + 20 Wednesday orders (stratified) – 40 Saturday orders answer Saturday only; the manager's remembered orders answer nothing. And the dictionary's 3 missing timer values are reported with the mean of the 37 complete cases, never silently.

2 Practice

2.1 For each field – delivery minutes; hour of day (11–21); rider code; satisfaction 1–5 – state quantitative or categorical, and the measurement scale. [4]

2.2 Nimbus wants to estimate average weekday delivery time to within ±1 minute. Rank these designs and justify: (a) every 10th order of the week, (b) a random sample of 200 weekday orders, (c) the 200 orders nearest the shop. [3]

2.3 The survey of 500 customers has a 22% response rate, and responders order 30% more often than average. Name the bias and state in one line what it does to the estimated satisfaction. [3]

2.4 A dataset has 40 rows and 3 missing delivery times. Two analysts report means of 27.43 (37 cases) and 25.52 (zeros filled). Which is defensible, and what must accompany it? [3]

exercises.lol exercises.com.cn

2.5 True or false, one line each: (a) a bigger convenience sample becomes representative; (b) stratified sampling needs the strata sizes known; (c) the population is every order the service could have taken that week. [3]

exercises.lol exercises.com.cn

Solutions

2.1

- Minutes: quantitative, ratio (zero = no time).
- Hour: categorical, ordinal.
- Rider code: categorical, nominal.
- Satisfaction: categorical, ordinal (ordered ratings).

2.2

- (b) random 200: unbiased, and n gives the ± 1.
- (a) systematic is near-random unless orders cycle in tens.
- (c) convenience: nearest orders are systematically faster – biased at any size.

2.3

- Non-response bias (with selection inside the response).
- Frequent orderers are over-represented.
- Satisfaction is biased toward loyal customers' view – state the direction, report the rate.

2.4

- The 37-case mean is defensible.
- Zero-filling invents three instant deliveries and drags the mean down by 1.9.
- The report must state the missingness and its likely direction (busy-runs forgotten $\rightarrow$ true mean likely a little higher).

2.5

- (a) False – size multiplies bias.
- (b) True – proportional allocation needs strata sizes (or estimates).
- (c) True – the population is defined by the claim you want to make.