

STATISTICS 1 CH1: DATA COLLECTION

Population: ALL of the possible data points / items of interest

Sample: Subset of population used to conduct a survey

Census: A sample that includes the entire population

Sampling unit: An item that can potentially be in a sample

Sampling frame: A list of all the sampling units

	Advantages	Disadvantages
Census	- Gives a completely accurate result	- Time consuming & expensive - Cannot be used in destructive tests - Lots of data to process
Sample	- Quicker & cheaper - Easier - Less data to process	- Data may not be accurate - Sample may not be large enough to represent small groups

When sample size ↑, result accuracy ↑

SAMPLING METHODS

Type	Name	Carrying Out	Advantages	Disadvantages
Every unit has equal chance	Simple Random Sampling	Assign # to units. Then, use random # generator to choose random sample	- Bias free - Easy & cheap - Each unit has a known chance	- Not suitable if pop. size ↑ - Sampling frame needed
	Systematic Sampling	Samples chosen at regular intervals (every k th item) $k = \frac{\text{pop size}(N)}{\text{sample size}(n)}$ start from random (1, k)	- Simple & quick - Suitable for large population	- Sampling frame needed - Can introduce bias if Sampling frame not random
	Stratified Sampling	Sample k% of pop., & pop. is divided into strata (ie groups), randomly sample k% of each strata	- Reflects pop. structure - Guarantees proportional representation of strata	- Pop. must be clearly classified - Selection within strata suffers from same disadvantages
Every unit doesn't have equal chance	Quota Sampling	Pop. divided into strata. Set quota for each stratum, then interviewer selects sampling units to fill quota	- Allows small sample size to still be representative - No sampling frame needed - Quick & easy - Comparison between groups	- Non-random = BIAS - Strata required - Scope ↑, time & expense ↑ - Non-response not recorded
	Opportunity Sampling	Find units at the same time as the survey is being carried out	- Easiest to carry out - Inexpensive	- Highly unlikely to provide representative sample - Highly dependent on interviewer

TYPES OF DATA

non-numerical value

Qualitative
e.g. colour

numerical value

Quantitative

Specific values

e.g. # of siblings

Discrete

Any decimal value

Continuous e.g. height

Both can be infinite or have range

You can "group results"

1, 2, 3, 4, 5, 6, 7, 8 $\leftarrow \uparrow$ precision, \downarrow conciseness



$1 \leq x \leq 5$ ####

$6 \leq x \leq 10$ / / /

$\leftarrow \downarrow$ precision, \uparrow conciseness

class interval

tally

lower class boundary

midpoint
8

upper class boundary

class width
 $10 - 6 = 4$

P.16 - 17 MIXED EX 1

1 a) $\frac{14.6 + 8.8 + 7.2 + 7.3 + 10.1}{5} = 9.6^\circ\text{C}$

b) I will generate 5 unique random numbers from 1-15 (inclusive), then match the numbers to the "day of month" in the table

c) Days: 5, 15, 12, 6, 2

$$\text{mean} = \frac{10.1 + 9.5 + 12.7 + 11.9 + 8.8}{5} = 10.6^\circ\text{C}$$

d) $\frac{14.6 + 8.8 + 7.2 + 7.3 + 10.1 + 11.9 + 12.2 + 12.1 + 15.2 + 11.1 + 10.6 + 12.7 + 8.9 + 10.0 + 9.5}{15}$

$$\approx 10.8^\circ\text{C}$$

c is less reliable \because sample

d is more reliable \therefore census

2 a) i census A: Accurately reflects population D: Time consuming

ii sample A: Quick and easy D: May not reflect the entire population

b) place elements in list and number them.

use a random # generator to generate 100 UNIQUE numbers

elements that correspond to the numbers generated are part of the sample

3 a) i population: a collection of all the possible sample units

ii sampling frame: a list of all the possible sample units

b) i a list of all licensed drivers

ii all residents of Oxford

4 a) Stratified: A: Can represent sub-groups D: Strata needed

b) Simple random: A: no bias D: Not suitable for large sample frame

5 a) The people who work in the offices are more likely to enjoy fine dining and have higher expectations

- b) i assign a number to every worker
starting from a random number between 1 and 10, he picks every 10th number and finds the workers with the matching numbers
- ii assign #1-100 to the office workers
assign #101-300 to the factory floor workers
randomly generate 10 unique numbers from range (1, 100)
randomly generate 20 unique numbers from range (101, 300)
interview the workers with the corresponding #'s
- iii go to the office and walk up to 10 random people to interview them.
go to the factory and walk up to 20 random people to interview them.

- 6 a) assign each student a # from 1 to 120
randomly generate 15 unique numbers.
Students with the numbers are part of the sample
- b) assign each girl a # from 1 to 64
assign each boy a # from 65 to 120
randomly generate 8 unique numbers from range 1 to 64
randomly generate 7 unique numbers from range 65 to 120
Students with the numbers are part of the sample

- 7 a) stratified sampling
b) this method is not biased. it also accurately represents small groups

- 8 a) opportunity Sampling
b) It is very easy and straightforward
c) continuous, because the weights can be any decimal value and is only limited by the precision of the apparatus used.

d) $\frac{70 + 76 + 82 + 74 + 78}{5} = 76$

e) $\frac{79 + 86 + 90 + 68 + 75}{5} = 79.6$

f) 2nd one, because it has springboks from multiple locations

g) collect more springbok from each location.

9 a) no because the number "1" was not chosen randomly and the other dates didn't have a chance to be chosen

b) Simple Random: A: the results will be truly random

D: we need to organize the dates into a list

Systematic: A: The process is quick and simple

D: we need to organize the dates into a list

c) continuous because the actual amount of rainfall can be any decimal value

d) $\frac{9.0 + 0.0 + 1.0 + 32.0 + 4.1 + 3.0}{6} = \frac{49}{6} \approx 8.2$

e) not reliable because not all months have been used in the sample and the sample size is very small