

Using statistical methods to make an inference

Exercise B: Sampling methods

- A researcher wishes to compare attributes of 24-month-old children who attend day-care centres in two different cities, A and B. As part of the research into city A, the researcher uses records from the day-care centre *TopTots*. To ensure anonymity in the report, each child in the study is assigned a number in the multivariate table of data. The variables being measured are weight, height, and distance from home to the day-care centre.

Number of child	Sex	Weight (kg)	Height (cm)	Distance (km)
1	Female	11.3	84	2.3
2	Female	8.6	79	1.3
3	Female	10.2	82	2.1
4	Female	11.7	85	1.7
5	Female	12.4	87	1.9
6	Female	10.8	83	3.1
7	Female	11.9	86	2.6
8	Female	10.6	83	3.3
9	Female	12.8	88	2.0
10	Female	11.9	86	1.5
11	Female	11.1	80	1.8
12	Female	11.5	85	2.1
13	Female	14.3	91	1.7
14	Female	13.4	89	2.3
15	Female	12.4	87	2.0
16	Female	14.7	92	2.9
17	Female	11.1	84	1.7
18	Female	12.0	86	2.6
19	Female	11.0	84	1.3
20	Female	13.5	89	2.4
21	Female	12.0	86	1.5
22	Female	11.6	85	2.1
23	Male	11.7	86	1.3
24	Male	13.0	89	3.0
25	Male	14.4	92	2.0
26	Male	10.9	84	2.4
27	Male	15.8	95	1.8
28	Male	13.0	89	2.3
29	Male	15.5	94	1.4
30	Male	11.8	86	1.1
31	Male	12.1	87	2.0
32	Male	13.5	90	2.5
33	Male	13.3	89	2.1

Number of child	Sex	Weight (kg)	Height (cm)	Distance (km)
34	Male	13.4	90	1.6
35	Male	12.2	87	2.0
36	Male	10.6	84	1.5
37	Male	13.9	91	1.7
38	Male	9.5	81	1.1
39	Male	12.2	87	1.5
40	Male	11.4	86	2.1
41	Male	12.8	89	1.1
42	Male	10.5	83	1.8
43	Male	13.0	89	2.4
44	Male	12.5	88	1.9
45	Male	11.1	85	1.2
46	Male	14.2	91	1.7
47	Male	13.3	90	2.0
48	Male	13.2	89	1.9
49	Male	14.0	91	1.5
50	Male	11.8	86	1.8
51	Male	14.2	91	2.0
52	Male	13.5	90	1.5
53	Male	14.4	92	1.7
54	Male	14.9	93	1.1
55	Male	13.8	91	2.1
56	Male	15.2	94	2.1
57	Male	13.5	90	2.0
58	Male	10.7	84	2.2
59	Male	13.6	90	1.4
60	Male	12.8	88	1.7
61	Male	12.6	88	1.4
62	Male	12.8	88	1.5
63	Male	14.2	91	2.3
64	Male	10.3	83	1.8
65	Male	13.5	90	2.0
66	Male	13.3	90	2.4

- A researcher wishes to select 12 random weights from the table using simple random sampling. Explain how this could be done. Show your selections by highlighting the resulting weights.

Using statistical methods to make an inference

- b. A random selection of 12 distances from home to the day-care centre is to be made using systematic sampling. Explain how this could be done. Show your selections by highlighting the resulting distances.

- c. The researcher considers that the sex of a child has an effect on his/her height, so a stratified sample of 12 heights is to be selected based on sex. Explain how this could be done. Show your selections by highlighting the resulting heights.

- d. The data from *TopTots* is combined with the data from seven other randomly selected day-care centres from city A to create the complete sample for city A which will be used in the investigation.

- i. What is this sampling method called?

- ii. Why would this method have been chosen? Give three reasons.

- iii. What is a potential problem with this method?

2. A factory has three machines producing 500 gram bags of rice. For this task, Machine A is used 30 hours per week, Machine B is used 20 hours per week, and Machine C is used 10 hours per week.

A quality-control supervisor takes a random sample of 120 bags weekly to check the actual weights of the bags. The table shows the actual weights (to the nearest whole gram) of one random sample of 120 bags.

Sample	1	2	3	4	5	6
1	505	503	500	501	504	498
2	505	507	505	511	500	506
3	501	501	503	502	501	504
4	503	503	505	506	503	503
5	504	502	501	503	505	497
6	504	501	503	510	504	502
7	502	502	503	499	506	501
8	501	495	507	504	500	509
9	503	500	503	503	502	498
10	505	501	507	504	501	508
11	503	504	501	504	505	500
12	500	504	503	508	502	505
13	498	501	497	505	501	504
14	506	506	498	504	508	509
15	504	500	510	503	511	502
16	504	506	503	505	509	504
17	504	507	502	501	504	494
18	503	503	506	506	507	503
19	503	502	512	503	512	515
20	500	504	507	510	499	507

- a. If the supervisor used stratified sampling, describe how this sample could have been selected. What assumptions would be made?

Using statistical methods to make an inference

b. If the supervisor used cluster sampling, describe how this sample could have been selected. What assumptions would be made?

iii. How many bags were selected twice (i.e. by both the simple random sample and the systematic sample)? Is this what you would have expected? What feature of sampling does this illustrate?

c. The supervisor decides to select 30 bags from the sample above for further testing.

i. Describe how you could use simple random sampling with pairs of numbers to select a sample of size 30 from the table above. Highlight the weights selected.

ii. Describe how you could use systematic sampling to select a sample of size 30 from the table above. Circle the weights selected.

3. Visit the CensusAtSchool website www.censusatschool.org.nz and use it to select a sample of 50 Year 12 students from your region of New Zealand. Comment on the process and any features of interest in the data in your sample.

Using statistical methods to make an inference

Answers

1. *Answers will vary.*
 - a. Enter $66 \text{ Ran\#} + 1$ into the calculator, then press = repeatedly. Take the whole-number part of each number produced, ignoring repeats, until a sample of 12 different numbers between 1 and 66 is produced. Highlight the weights corresponding to these numbers on the chart.
 - b. Step size = $66 \div 12 = 6$ (to nearest whole number). Random starting point selected within the first step using simple random sampling ($6 \text{ Ran\#} + 1 =$) to get 4, say. The 11 numbers chosen systematically are 4, 10, 16, ..., 64. The 12th number can be chosen randomly ($66 \text{ Ran\#} + 1$) ignoring any number already in the set of selected numbers. Highlight the distances corresponding to these numbers on the chart.
 - c. One third ($\frac{22}{66}$) of the population is female so one third of the sample should be female, i.e. 4 females. To choose these females, enter $22 \text{ Ran\#} + 1$ and press = repeatedly, taking the whole-number part and ignoring repeats until 4 different numbers are produced between 1 and 22. Highlight the heights corresponding to these numbers on the chart. Similarly, 8 males are chosen. Enter $44 \text{ Ran\#} + 23$ and press = repeatedly, taking the whole-number part and ignoring repeats until 8 different numbers are produced between 23 and 66. Highlight the heights corresponding to these numbers on the chart.
 - d.
 - i. Cluster sampling
 - ii. Less travel, time, cost
 - iii. Clusters may not be truly representative of population.
2. a. *Answers will vary.*

If systematic sampling is used, then the origin of the bag (which machine produced it) is assumed to be of importance (i.e. the machines have different characteristics from each other). If each machine produces bags per hour at the same rate, then Machine A produces $\frac{30}{60}$ or $\frac{1}{2}$ of the bags, Machine B produces $\frac{20}{60}$ or $\frac{1}{3}$ of the bags, Machine C produces $\frac{10}{60}$ or $\frac{1}{6}$ of the bags. So in the sample of 120 bags, 60 bags were from Machine A, 40 bags were from Machine B and 20 bags were from Machine C.
- b. If cluster sampling is used, then the bags may have all been produced by a single machine. In this case it is assumed that there are no differences between the outputs of the three machines.
- c. *Answers will vary.*
 - i. Each bag is in one of 20 rows, and one of 6 columns. Generate random numbers in pairs: the first from 1–20 inclusive ($20 \text{ Ran\#} + 1$) which gives the row, and the second from 1–6 inclusive ($6 \text{ Ran\#} + 1$) which gives the column. For each pair of numbers, a cell in the grid is selected. For example, if (row, column) = (17, 4) then the bag is in the 17th row and the 4th column and has weight 501 g.
 - ii. Step size = $120 \div 30 = 4$. Random starting number ($4 \text{ Ran\#} + 1 =$) gives 2, say. Start at the 2nd bag weight in the first column then highlight every 4th bag after that.
 - iii. The number of twice-selected bags will vary. This is expected because of the variability of sampling. There is an extremely large number of possible samples of 30 from 120. It is very unlikely any two samples are exactly the same.
3. *Answers will vary.*

Interesting features could include dirty data (inappropriate or missing data). It may also be interesting to see if the males and females each make up 50% of the sample. On average this would be expected, but due to sampling variability it is unlikely to be true for this sample.