



ANSWER BOOK

Sampling and the large data set

Statistics · Year 12–13

7 questions · 23 marks

NAVY EDITION

SECTION A · Warm-up

- A1** The population is every member of the group under study. A census measures all of them. A sample measures a chosen subset and uses it to say something about the whole. B1 B1 B1 for the three definitions. A census is complete but slow and expensive, and for destructive testing it is impossible. A sample trades completeness for practicality. [3]
- A2** You need a sampling frame, which is a numbered list of the whole population. Every member must then have an equal chance of selection, and every possible sample of that size must be equally likely. B1 for the sampling frame, B1 for the equal chance of selection. Random numbers or a lottery draw deliver that; picking whoever is nearest never does. [2]

SECTION B · Standard

- B1** Do the proportions first. First years supply $50 \times 720/1200 = 30$ students and second years $50 \times 480/1200 = 20$, and the two rebuild the sample size of 50. Then, within each year group, number the students and take a simple random sample of the required size using random numbers. M1 for the proportions, A1 for 30 and 20, B1 for numbering within each year group, B1 for the random selection from each. Both halves earn marks, and candidates who do the arithmetic and stop lose the second half every time. Stratifying is worth the trouble here because it guarantees the sample matches the college's 60:40 split, whereas a simple random sample matches it only on average and one unlucky draw could badly over-represent a year. [4]
- B2** The interval is $1200/60 = 20$. Choose a random starting point between 1 and 20, then take every 20th name from there. B1 for the interval, B1 for the random start, B1 for the rule. The random start is what keeps the method fair; always beginning at the first name would not be. State the interval, the random start and the rule, because all three are needed before the sample can be reproduced. [3]
- B3** Opportunity sampling, also called convenience sampling, since it takes whoever happens to be available. The people arriving at a gym at 7 am exercise far more than the population the survey is meant to describe, so every estimate drawn from them overstates activity levels in the same direction. B1 for naming the method, B1 for the group over-represented, B1 for the direction of the bias. Naming the over-represented group is what earns the explanation mark. Ease of collection is the method's only advantage. [3]

SECTION C · Stretch

- C1** First, Heathrow is one of only five UK stations in the data set, and one site in west London cannot speak for the whole country. Camborne and Leuchars have very different climates. Second, the claim is about a change over time, yet only 2015 has been used. Without the matching 1987 figures there is nothing to compare, and two summers 28 years apart would in any case be too little to establish a trend. A third reason worth having: each period runs only from May to October, so half of every year is missing and nothing follows about winter rainfall. As for the awkward entries, tr means a trace, a measurable fall below 0.05 mm, so it should be entered as zero or as a very small value rather than skipped. Entries of n/a are missing data and must be left out of the calculation altogether, which reduces the number of days the mean is divided by. B1 for the first reason, B1 for supporting it with the other stations, B1 for the second reason, B1 for supporting it with the missing earlier data, B1 for the treatment of tr and n/a. Silently treating either as zero, or as an ordinary reading, shifts the mean and is exactly the slip these questions are set to catch. [5]
- C2** The sampling frame excludes the very people the survey most needs, because anyone with poor or no internet access cannot reach the form. The responses will therefore overstate how reliable the internet is. No increase in sample size repairs this, since the missing group is missing systematically rather than by chance. B1 for the frame excluding those with poor access, B1 for the responses overstating reliability, B1 for a larger sample not helping. [3]
-