1. A large company surveyed its staff to investigate the awareness of company policy. The company employs 6000 full time staff and 4000 part time staff.

(a) Describe how a stratified sample of 200 staff could be taken.

(3)

(b) Explain an advantage of using a stratified sample rather than a simple random sample.

(1)

A random sample of 80 full time staff and an independent random sample of 80 part time staff were given a test of policy awareness. The results are summarised in the table below.

	Mean score ($\bar{x}$)	Variance of scores (s^2)
Full time staff	52	21
Part time staff	50	19

(c) Stating your hypotheses clearly, test, at the 1% level of significance, whether or not the mean policy awareness scores for full time and part time staff are different.

(7)

(d) Explain the significance of the Central Limit Theorem to the test in part (c).

(2)

(e) State an assumption you have made in carrying out the test in part (c).

(1)

After all the staff had completed a training course the 80 full time staff and the 80 part time staff were given another test of policy awareness. The value of the test statistic z was 2.53

(f) Comment on the awareness of company policy for the full time and part time staff in light of this result. Use a 1% level of significance.

(2)

(g) Interpret your answers to part (c) and part (f).

(1)
(Total 17 marks)

2. A telephone directory contains 50 000 names. A researcher wishes to select a systematic sample of 100 names from the directory.

(a) Explain in detail how the researcher should obtain such a sample.

(2)

(b) Give one advantage and one disadvantage of

(i) quota sampling,

(ii) systematic sampling.

(4)
(Total 6 marks)

3. A company produces climbing ropes. The lengths of the climbing ropes are normally distributed. A random sample of 5 ropes is taken and the length, in metres, of each rope is measured. The results are given below.

120.3 120.1 120.4 120.2 119.9

(a) Calculate unbiased estimates for the mean and the variance of the lengths of the climbing ropes produced by the company.

(5)

The lengths of climbing rope are known to have a standard deviation of 0.2 m. The company wants to make sure that there is a probability of at least 0.90 that the estimate of the population mean, based on a random sample size of n, lies within 0.05 m of its true value.

(b) Find the minimum sample size required.

(6)
(Total 11 marks)

4. A researcher is hired by a cleaning company to survey the opinions of employees on a proposed pension scheme. The company employs 55 managers and 495 cleaners.

To collect data the researcher decides to give a questionnaire to the first 50 cleaners to leave at the end of the day.

(a) Give 2 reasons why this method is likely to produce biased results.

(2)

(b) Explain briefly how the researcher could select a sample of 50 employees using

(i) a systematic sample,

(ii) a stratified sample.

(6)

Using the random number tables in the formulae book, and starting with the top left hand corner (8) and working across, 50 random numbers between 1 and 550 inclusive were selected. The first two suitable numbers are 384 and 100.

(c) Find the next two suitable numbers.

(2)
(Total 10 marks)

5. Describe one advantage and one disadvantage of

(a) quota sampling,

(2)

(b) simple random sampling.

(2)
(Total 4 marks)

6. A school has 15 classes and a sixth form. In each class there are 30 students. In the sixth form there are 150 students. There are equal numbers of boys and girls in each class. There are equal numbers of boys and girls in the sixth form. The head teacher wishes to obtain the opinions of the students about school uniforms.

Explain how the head teacher would take a stratified sample of size 40.

(Total 7 marks)

7. (a) State two reasons why stratified sampling might be chosen as a method of sampling when carrying out a statistical survey.

(2)

(b) State one advantage and one disadvantage of quota sampling.

(2)
(Total 4 marks)

8. There are 64 girls and 56 boys in a school.

Explain briefly how you could take a random sample of 15 pupils using

(a) a simple random sample,

(3)

(b) a stratified sample.

(3)
(Total 6 marks)

9. Explain briefly what you understand by

(a) a statistic,

(2)

(b) a sampling distribution.

(2)
(Total 4 marks)

10. Explain how to obtain a sample from a population using

(a) stratified sampling,

(2)

(b) quota sampling.

(2)

Give one advantage and one disadvantage of each sampling method.

(4)
(Total 8 marks)

11. A bag contains a large number of coins of which 30% are 50p coins, 20% are 10p coins and the rest are 2p coins.

(a) Find the mean μ and the variance σ^2 of this population of coins.

(4)

A random sample of 2 coins is drawn from the bag one after the other.

(b) List all possible samples that could be drawn.

(2)

(c) Find the sampling distribution of $\overline{X}$, the mean of the coins drawn.

(4)

(d) Find $P(2 \leq \overline{X} < 7)$.

(2)

(e) Use the sampling distribution of $\overline{X}$ to verify $E(\overline{X}) = \mu$ and $\text{Var}(\overline{X}) = \frac{1}{2}\sigma^2$.

(5)
(Total 17 marks)

1.	(a)	Label full time staff 1 – 6000, part time staff 1 – 4000		
		Use random numbers to select		
		Simple random sample of 120 full time staff and 80 part time staff	A1	3

Note

1st for attempt at labelling full-time and part-time staff. One set of correct numbers.

2nd for mentioning use of random numbers

1st A1 for s.r.s. of 120 full-time and 80 part-time

(b)	Enables estimation of statistics / errors for each strata <u>or</u> "reduce variability" <u>or</u> "more representative" <u>or</u> "reflects population structure" **NOT** "more accurate"	B1	1

(c) H_0: $\mu_f = \mu_p$, H_1: $\mu_f \neq \mu_p$ (accept μ_1, μ_2) B1

s.e. = $\sqrt{\frac{21}{80}+\frac{19}{80}}$, $z = \frac{52-50}{\sqrt{\frac{21}{80}+\frac{19}{80}}} = \left(2\sqrt{2}\right)$

= 2.828… (awrt **2.83**) A1

Two tailed critical value z = 2.5758 (or prob of awrt 0.002 (<0.005) or 0.004 (<0.01)) [2.828 > 2.5758 so] significant evidence to reject H_0 B1 dM1

There is evidence of a difference in policy awareness between full time and part time staff A1ft 7

Note

1st for attempt at s.e. – condone one number wrong. NB correct s.e. = $\sqrt{\frac{1}{2}}$

2nd for using their s.e. in correct formula for test statistic. Must be $\frac{\pm(52-50)}{\sqrt{\frac{p}{q}+\frac{r}{s}}}$

3rd dM1 **dep. on 2nd** for a correct statement based on their normal cv and their test statistic

2nd A1 for correct comment in context. Must mention "scores" or "policy awareness" and types of "staff". Award **A0** for a one-tailed comment. Allow ft

(d)	Can use mean full time and mean part time	B1	
	~ Normal	B1	2

Note

1st B1 for mention of mean(s) <u>or</u> use of $\bar{X}$, provided $\bar{X}$ clearly refers to full-time or part-time

		2^{nd} B1 for stating that distribution can be assumed normal e.g. "mean score of the test is normally distributed" gets B1B1		
	(e)	Have assumed $s^2 = \sigma^2$ or variance of sample = variance of population	B1	1
	(f)	2.53 < 2.5758, not significant <u>or</u> do not reject H_0 So there is insufficient evidence of a difference in mean awareness	A1ft	2

Note

for correct statement (may be implied by correct contextualised comment)

A1 for correct contextualised comment. Accept "no difference in mean scores". Allow ft

	(g)	Training course has closed the gap between full time staff and part time staff's mean awareness of company policy.	B1	1

Note

B1 for correct comment in context that implies training was effective. This must be supported by their (c) and (f). Condone one-tailed comment here.

[17]

2.	(a)	Randomly select a number between 00 and 499 (001 and 500)	B1	
		select every 500^{th} person	B1	2

Note

1^{st} B1 for idea of using random numbers to select the first from1 – 500 (o.e.)

2^{nd} B1 for selecting every 500^{th} (name on the list)

If they are clearly trying to carry out <u>stratified</u> sample then score B0B0

(b)	(i)	Quota **Adantage:** Representative sample can be achieved (with small sample size) Cheap (costs kept to a minimum) **not** "quick" Administration relatively easy	B1	
		Disadvantage Not possible to estimate sampling errors (due to lack of randomness) Not a random process Judgment of interviewer can affect choice of sample – bias Non-response not recorded Difficulties of defining controls e.g. social class	B1	2
		Note Score B1 for any one line 1^{st} B1 for Quota advantage 2^{nd} B1 for Quota disadvantage		
	(ii)	Systematic **Advantage:** Simple or easy to use **not** "quick" or "cheap" or "efficient" It is suitable for large samples (not populations)	B1	
		Disadvantage Only random if the ordered list is (truly) random Requires a list of the population or must assign a number to each member of the pop.	B1	2
		Note Score B1 for any one line 1^{st} B1 for Systematic Advantage 2^{nd} B1 for Systematic Disadvantage		

[6]

3. (a) Estimate of Mean = $\frac{600.9}{5}$ =120.18 M1A1

Estimate of Variance = $\frac{1}{4}\{72216.31 - \frac{600.9^2}{5}\}$

or $\frac{0.148}{4} = 0.037$

A1ft A1 5

<u>Note</u>

1st	for an attempt at Σx (accept 600 to 1sf)
1st A1	for $\frac{600.9}{5}$ = awrt 120 or awrt 120.2. No working give M1A1 for awrt 120.2
2nd	for the use of a correct formula including a reasonable attempt at $\sum x^2$ (Accept 70 000 to 1sf) or $\sum\left(x-\bar{x}\right)^2$ =0.15 (to 2 dp)
2nd A1ft	for a correct expression with correct $\sum x^2$ but can ft their <u>mean</u> (for expression – no need to check values if it is incorrect)
3rd A1	for 0.037 Correct answer with no working scores 3/3 for variance

(b) P(–0.05 < $\mu - \hat{\mu}$ < 0.05) = 0.90 or P(–0.05< $\bar{X} - \mu$ <0.05) =0.90 [$\le$ is OK] B1

$$\frac{0.05}{\frac{0.2}{\sqrt{n}}} = 1.6449$$ A1

$$n = \frac{1.6449^2 \times 0.2^2}{0.05^2}$$ dM1

n = 43.29… A1

n = 44 A1 6

Note

B1 for a correct probability statement or "width of 90% CI = 0.05×2 = 0.1" — 1st B1 may be implied by 1st A1 scored or correct equation.

1st for $\frac{0.05}{\frac{0.2}{\sqrt{n}}} = z$ value or $2 \times \frac{0.2}{\sqrt{n}} \times z = 0.1$

Condone 0.5 instead of 0.05 or missing 2 or 0.05 for 0.1 for

1st A1 for a correct equation including 1.6449

2nd dM1 Dependent upon 1st for rearranging to get n = …Must see "squaring"

2nd A1 for n = awrt 43.3

3rd A1 for rounding up to get n = 44

Using e.g.1.645 instead of 1.6449 can score all the marks except the 1st A1

[11]

4. (a) Only cleaners – no managers i.e. not all types. OR Not a random sample B1g
1st 50 may be in same shift/group/share same views.
OR Not a random sample B1h 2
(Allow "not a representative sample" in place of "not a random sample")

After 1st B1, comments should be in **context**, i.e. mention cleaners, managers, types of worker etc

1st B1g for one row

2nd B1h for both rows. "Not a random sample" only counts once.
Score B1B0 or B1B1 or B0B0 on EPEN

(b) (i) Label employees (1–550) or obtain an ordered list B1
Select first using random numbers (from 1 – 11) B1
Then select every 11th person from the list B1

1st B1 for idea of labelling or getting an ordered list. No need to see 1–550.

2nd B1 selecting first member of sample using random numbers (1–11 need not be mentioned)

3rd B1 selecting every nth where n = 11.

(ii) Label managers (1–55) and cleaners (1–495)
Use random numbers to select…
…5 managers and 45 cleaners — A1 — 6

(c) 390, 372 (They must be in this order) — B1, B1 — 2

1st for idea of two groups and labelling both groups.
(Actual numbers used not required)

2nd for use of random numbers within each strata.
Don't give for SRS from all 550.
"Assign random numbers to managers and cleaners" scores M0M1

A1 for 5 managers and 45 cleaners.
(This mark is dependent upon scoring at least one M)

[10]

5. (a) Advantages:

- does not require the existence of:

  a sampling frame
  a population list

- field work can be done quickly as representative sample can be achieved with a small sample size
- costs kept to a minimum (cheaply)
- administration relatively easy
- non-response not an issue — B1

any one

Disadvantages:

- not possible to estimate sampling errors
- interviewer choice and may not be able to judge easily / may lead to bias
- non-response not recorded
- non-random process — B1 — 2

any one

(b) Advantages:

- random process so possible to estimate sampling errors
- free from bias — B1

any one

Disadvantages:

- not suitable when sample size is large
- sampling frame required which may not exist or may be difficult to construct for a large population — B1 — 2

any one

NO REPETITION / OPPOSITES

[4]

6. Total in School = (15 × 30) + 150 = 600 — B1

random sample of $\frac{30}{600} \times 40$ (Use of $\frac{40}{their\,600}$)

= 2 from each of the 15 classes — A1

random sample of $\frac{150}{600} \times 40$ Either

= 10 from sixth form; — A1

Label the boys in each class from 1 – 15 and the girls from 1 – 15. — B1
use random numbers to select 1 girl and 1 boy — B1

Label the boys in the sixth form from 1 – 75 and the girls from 1 – 75. — B1
use random numbers to select 5 different boys and 5 different girls.

[7]

7. (a) Population divides into mutually exclusive; groups — B1; B1 — 2
distinct strata

(b) Advantages
– enables fieldwork to be done quickly
– costs kept to a minimum
– administration is relatively easy — B1

Any one

Disadvantages
– non-random so not possible to estimate sampling errors
– subject to possible interviewer bias — B1 — 2
– non-response not recorded

Any one

[4]

8.	(a)	Allocate a number between 1 and N (or equiv) to each pupil.			
		Use random number tables, computer or calculator to select 15 different numbers between 1 and 120 (or equiv).		B1	
		Pupils corresponding to these numbers become the sample.		B1	3
	(b)	Allocate numbers 1 – 64 to girls and 1 – 56 to boys. Idea of different sets for boys and girls			
		Select $\frac{64}{120} \times 15 = 8$ random numbers between 1 – 64 for girls			
		attempt find no			
		Select 7 random numbers between 1 – 56 for boys.		A1	3
		Both 7 and 8			
					[6]
9.	(a)	A random variable; that is, a function involving no unknown quantities		B1; B1	2
	(b)	If all possible samples are taken; then their values will form a distribution called the sampling distribution		B1; B1	2
					[4]
10.	(a)	Take a (simple) random sample from (mutually exclusive) groups of the population	1g/1h	B1	
		Sample sizes within strata in strict proportion to numbers in each strata in the population		B1	
		Advantage: More accurate estimate of variance of population mean Individual estimates for strata available	Any one	B1	
		Disadvantage: Difficult if strata are large Definition of strata problematic (may overlap)	Any one	B1	4

(b) Non-random sampling — B1
from groups of the population — B1 dep

Advantage:
Representative sample can be achieved with small sample size
Cheap (costs kept to a minimum)
Administration relatively easy — Any one (not quick) — B1

Disadvantage
Not possible to estimate sampling errors due to lack of randomness
Judgment of interviewer can affect choice of sample – bias OK
Non-response not recorded
Difficulties of defining controls e.g. social class — Any one — B1 — 4

[8]

11. (a) μ 0.3 × 50 + 0.2 × 10 + 0.5 × 2 = 18 — A1
$\sigma^2 = (0.3 \times 50^2 + 0.2 \times 10^2 + 0.5 \times 2^2) - 18^2 = 448$ — A1 — 4

(b)

	or	
(50,50)		(50,50) without ordered pairs
(10,2)		(10,2)
(2,10)		(10,10)
(10,10)		(50,10)
(50,10)		(2,2)
(10,50)		(50,2)
(2,2)		
(50,2)		
(2,50)		

either, –1 each missing pair — B2 — 2

(c)

$\bar{x}$	2	6	10	26	30	50
$P(\bar{X} = \bar{x})$	0.25	0.2	0.04	0.3	0.12	0.09

All means, probabs multiplied, –1 each error — B1 — A2 — 4

(d) $P(2 \le \bar{X} < 7) = 0.25 + 0.2 = 0.45$ — A1∫ — 2
Probabilities of 2 and 6 added, 0.45

$E(\bar{X}) = 2 \times 0.25 + 6 \times 0.2 \ldots = 18$ — $\Sigma x P(X = x)$ from table, 18 — A1
$\text{Var}(\bar{X}) = 2^2 \times 0.25 + 6^2 \times 0.2 + \ldots - 18^2 = 224$ — A1
$\Sigma x^2 P(X = x) - (\text{theirs})^2$, 224

So $E(\bar{X}) = 18 = \mu$ and $\text{Var}(\bar{X}) = 224 = \frac{1}{2}\sigma^2$ as required. — A1 — 5

[17]

1. Most candidates knew how to take a stratified sample by taking simple random samples in each stratum but they often forgot to describe how to label the members of the strata.

In (b) the commonest correct response was about the sample being more representative of the population but some missed the point and simply said that stratified sampling was “easier”.

The calculation in part (c) was carried out very well by most candidates. There were few errors with the standard error and most correctly concluded that there was evidence of a difference in policy awareness between the types of staff..

In part (d) most knew that the Central Limit Theorem had something to do with the normal distribution but they did not mention that it was the mean scores of full time and part time staff that can be assumed to be normally distributed.

There were some correct responses to part (e) but many just mentioned independence despite this being given in the stem to part (c) of the question.

Most gave a correct conclusion in part (f) and some correctly inferred in the final part that the training course had been effective.

Some had the correct idea in part (g) although their conclusions went further than the evidence suggested: they claimed that the scores of the part time staff had increased, which may well be the case, but the evidence presented was only sufficient to conclude that the “gap” between policy awareness of the types of staff has been closed.

2. Most candidates realised that they would need to sample every 500^{th} name on the list in part (a) but a number did not explain how to select the first member of their sample at random from the first 500 names. In part (b) it was clear that many candidates had learnt some standard reasons from a textbook and nearly everyone scored something here. A few candidates confused systematic and stratified sampling in (ii) but for the most part this question was answered quite well.

3. Most answered part (a) well with the calculations being clearly laid out. Part (b) caused many problems. Many only considered one “tail” so they effectively used P($\overline{X} - \mu < 0.05$) = 0.90, thers realised that the calculations from a confidence interval were involved and equations involving $\frac{0.2}{\sqrt{n}} \times z$ appeared but they were not always correct. Some did muddle through to the correct answer but there were few clearly set out solutions.

4. Questions of this type are usually quite challenging for candidates and examiners alike. Candidates should be aware of how many marks the examiner is seeking to award in each part of the question and try and ensure that they make that many independent points. The use of bullet points rather than continuous prose might help both candidates and examiners.

In part (a) a number of candidates missed the fact that the sample contained no managers. Most knew how to take a systematic sample and explained the need to label the employees and pick every 11^{th} one but a mechanism for selecting the first one at random was often not mentioned. The stratified sampling procedure was well known too and usually applied to this situation quite well. Some lost a mark for failing to label the managers and cleaners or for not using random numbers when selecting the samples from each strata. Part (c) proved an easy two marks for those who knew how to use the random number tables and most did know!

5. In part (a) they usually scored well, but there were few completely correct answers. Many candidates had little to offer but regurgitated text book definitions.

6. Most candidates calculated that there were 600 students in the school and they used classes and gender as strata. The calculations to determine the number of boys and girls from each class and the sixth form were often carried out correctly although sometimes they forgot there were 15 classes and simply suggested that a sample of 15 boys and 15 girls was taken rather than one boy and one girl from each class. The commonest omission was a failure to explain how the samples were taken from each stratum: labelling and using random numbers. There were a good number of fully or almost fully correct solutions from candidates who appreciated the depth of explanation required for a 7 mark question.

7. Many candidates were unaware of the reasons for the use of stratified sampling but most could give one advantage and one disadvantage of quota sampling.

8. Full marks were rarely gained on this question. In part (a) while most candidates realised that they needed to allocate numbers and use some form of random number generator many did not mention the need to choose 15 different numbers and use the pupils corresponding to these numbers. In part (b) candidates usually failed to allocate different sets of numbers to the boys and the girls. Many managed to gain two marks by working out that samples of 8 girls and 7 boys were needed.

9. Very few candidates scored full marks on this question. It would appear that many of them had not learnt the definitions. A common answer to part (a) was the definition of a statistical model and in part (b) the candidates confused sampling distributions with a sampling frame or with the main distributions such as a Binomial.

10. The definition of stratified sampling was attempted well, but candidates usually missed 'non-random' in the definition of quota sampling. Too many answers were vague and features were confused with advantages / disadvantages.

11. The first two parts were done well, with many gaining 6 marks. Thereafter it was clear that candidates had not focussed on this area of the syllabus. Candidates had great difficulty generating the correct means and many did not even try. Probabilities were achieved with varying success, but even strong candidates confused themselves by finding the expectation and variance of a sum of random variables. The candidates who wrote down the correct sampling distribution often went on to get full marks.