1.   (a)   A population is collection of all items   B1   1

   Note

   B1 – collection/group **al**l items – need to have /imply all eg
   entire/complete/every

   (b)   (A random variable) that is a function of the sample which contains   B1   1
         no unknown quantities/parameters.

   Note

   B1 – needs <u>function/calculation</u>(o.e.) <u>of the sample/random</u>
   <u>variables/observations</u> **and** <u>no unknown quantities/parameters</u>(o.e.)
   NB do not allow unknown variables
   e.g. "A calculation based <u>solely</u> on observations from a given sample." B1
       "A calculation based <u>only</u> on known data from a sample" B1
       "A calculation based on known observations from a sample" B0

   Solely/only imply no unknown quantities

   (c)   The voters in the town   B1

         Percentage/proportion voting for Dr Smith   B1   2

   Note

   B1 – Voters

   Do not allow 100 voters.

   B1 – percentage/ proportion voting (for Dr Smith)
           the **number** of people voting (for Dr Smith)
           Allow 35% of people voting (for Dr Smith)
           Allow 35 people voting (for Dr Smith)
           Do **not** allow 35% or 35 alone

   (d)   Probability Distribution of those voting for Dr Smith from all   B1   1
         possible samples (of size 100)

   Note

   B1 – answers must include all three of these features
           (i)   All possible samples,
           (ii)   their associated probabilities,
           (iii) context of voting for Dr Smith.

   e.g "It is all possible values of the percentage and their associated
   probabilities." B0 no context

                                                                                        [5]

2. (a)

x	1p	2p
P($X = x$)	$\dfrac{1}{4}$	$\dfrac{3}{4}$

$$\mu = 1 \times \frac{1}{4} + 2 \times \frac{3}{4} = \frac{7}{4} \text{ or } 1\frac{3}{4} \text{ or } 1.75 \qquad\qquad \text{B1}$$

$$\sigma^2 \quad = 1^2 \times \frac{1}{4} + 2^2 \times \frac{3}{4} - \left(\frac{7}{4}\right)^2 \qquad\qquad \text{M1}$$

$$= \frac{3}{16} \text{ or } 0.1875 \qquad\qquad \text{A1} \qquad 3$$

<u>Note</u>

B1 1.75 oe

M1 for using $\sum\left(x^2 p\right) - \mu^2$

A1 0.1875 oe

(b) (1,1,1), (1,1,2) any order, (1,2,2) any order, (2,2,2)   B1

   (1,2,1) (2,1,1) (2,1,2) (2,2,1)   all 8 cases considered.   B1   2

                                         May be implied by 3 *
                                         (1,1,2) and 3 * (1,2,2)

<u>Note</u>

ignore repeats

(c)

$\bar{x}$	1	$\dfrac{4}{3}$	$\dfrac{5}{3}$	2
P($\bar{X} = \bar{x}$)	$\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} = \frac{1}{64}$	$3 \times \frac{1}{4} \times \frac{1}{4} \times \frac{3}{4} = \frac{9}{64}$	$3 \times \frac{1}{4} \times \frac{3}{4} \times \frac{3}{4} = \frac{27}{64}$	$\frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} = \frac{27}{64}$

                                           B1 M1 A1 M1 A1A1   6

Note

1st B1 4 correct means (allow repeats)

1st M1 for p^3 for either of the ends

1st A1 for 1/64or awrt 0.016 **and** 27/64 or awrt 0.422

2nd M1 $3 \times p^2(1 – p)$for either of the middle two $0 < p < 1$

May be awarded for finding the probability of the 3 samples with mean of either 4/3 or 5/3 .

2nd A1 for 9/64 (or 3/64 three times) and 27/64 (or 9/64 three times) accept awrt 3dp.

3rd A1 fully correct table, accept awrt 3dp.

[11]

3. (a) *A statistic* is a function of $X_1, X_2, ... X_n$ B1

that does not contain any unknown parameters   B1   2

Note

Examples of other acceptable wording:

B1 e.g. is a function of the sample or the data / is a quantity calculated from the sample or the data / is a random variable calculated from the sample or the data

B1 e.g. does not contain any unknown parameters/quantities contains only known parameters/quantities <u>only</u> contains values of the sample

Y is a function of $X_1, X_2, ... X_n$ that does not contain any unknown parameters	B1B1
is a function of the values of a sample with no unknowns	B1B1
is a function of the sample values	B1B0
is a function of all the data values	B1B0
A random variable calculated from the sample	B1B0
A random variable consisting of any function	B0B0
A function of a value of the sample	B1B0
A function of the sample which contains no other values/ parameters	B1B0

(b)   The <u>probability</u> distribution of *Y* or the
      distribution of all possible values of *Y* (o.e.)   B1   1

      <u>Note</u>

      Examples of other acceptable wording

      All possible values of the statistic together
      with their associated probabilities

(c)   Identify (ii) as not a statistic   B1

      Since <u>it contains</u> unknown parameters <u>μ and σ</u> .   dB1   2

      <u>Note</u>

      1stB1 for selecting only (ii)

      2nd B1 for a reason. This is dependent upon the
      first B1. Need to mention at least one of mu (mean)
      or sigma (standard deviation or variance) or
      unknown parameters.

      Examples

      since it contains mu B1

      since it contains sigma B1

      since it contains unknown parameters/quantities B1

      since it contains unknowns B0

                                                                   [5]

4.   (a)   A census is when <u>every member</u> of the <u>population</u> is investigated.   B1

         B1 Need one word from each group
         (1) <u>Every member /all items / entire</u> /oe
         (2) <u>population/collection of individuals/sampling frame</u>/oe

         enumerating the population on its own gets B0

   (b)   There would be no cookers left to sell.   B1

         B1 Idea of Tests to destruction. Do not accept cheap or quick

   (c)   A list of the unique identification numbers of the cookers.   B1

         B1 Idea of list/ register/database of cookers/serial numbers

(d)   A cooker   B1   4

       B1 cooker(s) / serial number(s)

       The sample of 5 cookers or every 400th cooker gets B1

                                                                                                                                             [4]

5.   (a)   Saves time / cheaper / easier   B1   1
                              any one

        or

        <u>A census / asking all members</u> takes a long time or is
        expensive or difficult to carry out

    (b)   <u>List, register or database of all club members / golfers</u>   B1   1

        or
        <u>Full membership list</u>

    (c)   Club <u>member(s)</u>   B1   1
                                                                                                                                  [3]

6.   (a)

X	1	2	5
P($X = x$)	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

        Mean = $1 \times \frac{1}{2} + 2 \times \frac{1}{3} + 5 \times \frac{1}{6}$ = 2 or 0.02 $\Sigma x.p(x)$ need ½ and $\frac{1}{3}$ M1 A1

        Variance = $1^2 \times \frac{1}{2} + 2^2 \times \frac{1}{3} + 5^2 \times \frac{1}{6} - 2^2 = 2$ or 0.0002   M A1   4

    (b) $\Sigma x^2 .p(x) - \lambda^2$
           (1,1)
           (1,2) and (2,1)   B2
           (1,5) and (5,1)   LHS   -1   B1   3

        e.e.

           (2,2)
           (2,5) and (5,2)   repeat of "theirs" on RHS   B1
           (5,5)

(c)

$\overline{x}$	1	1.5	2	3	3.5	5
$P(\overline{X} = \overline{x})$	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$	$\frac{1}{6}$	$2 \times \frac{1}{3} \times \frac{1}{6} = \frac{1}{9}$	$\frac{1}{36}$

                                 ¼   M1A1

                        1.5+,-1ee   M1 M2   6

                                                          [13]

7. (a) Individual member or element of the population or sampling frame   B1   1

(b) A <u>list</u> of <u>all</u> sampling units or <u>all</u> the population   B1   1

(c) <u>All</u> possible <u>samples</u> are chosen from a population; the <u>values</u> of   B1 B1   2
a <u>statistic</u> and the associated <u>probabilities</u> is a sampling distribution

                                                             [4]

8. (a) (i) A <u>collection</u> of individuals or <u>items</u>   B1

     (ii) A <u>list</u> of <u>all sampling units</u> in the population   B1   2

(b) Not always possible to keep this list up to date   B1   1

(c) (i) eg:- Pupils in year 12 – small easily listed population   B1
         <u>Population known</u> & <u>easily accessed</u>   B1

     (ii) Students in a University – large not easily listed population   B1
         <u>Population known</u> but too <u>time consuming/expensive</u>
         to interview all of them   B1   4

                                                             [7]

OR

(c) (i) Definition of census by example   B1

     (ii) Definition of sample by example   B1

9. (a) A <u>list of</u> (all) the members of the <u>population</u>   B1   1
    A random variable that is a <u>function</u> of a random <u>sample</u>   B1
    that contains <u>no unknown parameters</u>   B1   2

                                                             [3]

10. (a) List of patients registered with the practice.
Require 'list' or 'register' or database or similar   B1   1

     (b) The patient(s)   B1   1

     (c) Adv: Quicker, cheaper, easier, used when testing results in
                destruction of item, quality of info about each sampling
                unit is often better.   Any one   B1

         Disadv: Uncertainty due to natural variation, uncertainty due
                 to bias, possible bias as sampling frame incomplete,
                 bias due to subjective choice of sample, bias due to
                 non-response   Any one   B1   2

     (d) Non-response due to patients registered with the practice but
who have left the area   B1   1

                                                                                            [5]

11. (a) All subscribers to the magazine   B1   1

     (b) A list of all members that had paid their subscriptions   B1   1

     (c) Members who have paid   B1   1

     (d) Advantage: total accuracy   B1
         Disadvantage: time consuming to obtain data and analyse it   B1   2

     (e) Let X represent the number agreeing to change the name
         $\therefore X \sim B(25, 0.4)$ B1
         $P(X = 10) = P(X \le 10) - P(X \le 9) = 0.1612$ M1 A1   3

     (f) $H_0: p = 0.40$, $H_1: p < 0.40$ B1, B1
         $P(X \le 6) = 0.0736 > 0.05 \Rightarrow$ not significant   M1 A1
         No reason to reject H_0 and conclude % is less than the editor believes   A1   5

     (g) Let X represent the number agreeing to change the name
         $\therefore X \sim B(200, 0.4)$
         $P(71 \le X < 83) \approx P(70.5 \le Y < 82.5)$ where $Y \sim N(80, 48)$ B1 B1

$$\approx P\left(\frac{70.5 - 80}{\sqrt{48}} \le X < \frac{82.5 - 80}{\sqrt{48}}\right)$$ M1 M1

         $\approx P(-1.37 \le X < 0.36)$ A1 A1
         $= 0.5533$ A1   7

                                                                            [20]

12. (a) *D* is continuous   B1   1

     (b) Sampling Frame is the list of competitors or their results,   B1

          e.g. label the results 1—200 and randomly select 36 of them   B1   2

     (c) *X* = no. of competitors with *A* = 2   *X* ~ B(36, $\frac{1}{3}$)

          X ≈ ~N(12, 8)   M1 A1

          $P(X \geq 20) \approx P\left(Z \geq \dfrac{19.5 - 12}{\sqrt{8}} \right)$ $\pm \frac{1}{2}$, '*z*'   M1 M1

          = P(*Z* ≥ 2.65…)   A1

          = 1 – 0.9960   = 0.004   A1   6

     (d) Probability is very low, so assumption of P(*A* = 2) = $\frac{1}{3}$ is unlikely.   B1 B1   2
          (Suggests P(*A* = 2) might be higher.)

                                                                                      [11]

1.   This was poorly done with very few candidates scoring full marks. Those candidates who had
         learnt standard definitions fared better than those who used their own understanding of the
         terms because they were less likely to leave out vital elements of the definitions. Even those
         who answered parts (a) and (b) correctly were then unable to apply these definitions in context.

         In part (a) a large majority of candidates omitted to mention "all", or its equivalent.

         Part (b) was well answered because many candidates used a standard definition. The most
         common errors were using "population" instead of "sample and omitting "no unknown
         parameters".

         In part (c) a substantial number of candidates were confused about "the population in this case".
         Many thought it to be the sample of 100 voters. Others were closer to the truth with "all the
         residents of the town", but did not earn the mark because they had failed to distinguish between
         registered voters and residents. The statistic was more easily identified.

         Part (d) was poorly answered with many candidates having no idea what a sampling distribution
         was and those that did being unable to put it into context. The sampling distribution of a
         proportion is arguably one of the hardest to get a grip on and articulate convincingly.

2.   A high proportion of candidates attempted the first two parts of this question successfully, with
         the majority of candidates getting at least one mark for part (b). Those less successful in part (a)
         either misread the question and ended up with a denominator of 3 for the probabilities or

         confused formulae for calculating the mean and variance and used, for example, $\sum \dfrac{xp(x)}{n}$ for

         the mean or used $E(X^2)$ for σ^2. The solution to part (c) proved beyond the capability of a
         minority of candidates but, for the majority, many exemplary answers were evident, reflecting
         sound preparation on this topic. Candidates who found all 8 cases in (b) usually gained four
         marks in part (c) for calculating the probabilities. For a small percentage of those candidates,
         calculating the means was difficult and hence completing the table correctly was not possible. A
         few candidates tried unsuccessfully to use the binomial to answer part (c).

3.   This question was either answered very well with some text book solutions, although it seemed
         that only a minority of candidates earned all five marks, or badly with some strange
         descriptions. A reasonable number of candidates responded with comments that were very close
         to those in the mark scheme: evidence possibly of deliberate preparation and learning whilst
         others had internalised the concepts and provided responses in their own words. Whilst these
         responses might not have matched the 'official' answers, they nevertheless captured the essence
         of the concepts and were fully acceptable. There was confusion with the definition of statistics
         and parameters and part (b) was often attempted badly with candidates not knowing the
         definition of a probability distribution. On the whole this was one of the worst answered
         questions in the paper.

         In part (a) candidates gave various definitions sometimes all muddled up. Not many candidates
         gave clear definitions but a common error was candidates writing "**any** *function*" or "*no* **other**
         quantities".

         In part (b) again the candidates had mixed success. A significant minority scored marks by
         knowing that a sampling distribution involved all possible values of the statistic and their
         associated probabilities.

         In part (c) many could identify (ii) correctly and a variety of reasons were seen. This part
         seemed to be done well even by candidates who could not answer part (a). It was interesting to

see that a relatively large proportion of candidates who earned both marks for part (c), were unable to achieve either of the two marks in part (a). There was a connection between parts (a) and (c) that many candidates failed to recognise. If those candidates who wrote "(ii) is not a statistic because it has unknown parameters" had then reflected on their responses to parts (a) and (c), they could then have gone back to modify their answer to (a) in order to earn more marks.

4.   Nearly all candidates achieved at least one of the available marks but it was disappointing that there were not more attaining full marks.

   (a)   Too many candidates referred to the national census rather than a general definition. Some felt an enumeration was adequate and others failed to recognise that EVERY member had to be investigated.

   (b)   A failure to put the question in context and consider the consequences of testing every item meant that some candidates scored 0 in this part of the question. A few candidates did not read the question carefully and used cheap and quick as their reasons why a census should not be used when the question specifically said give a reason "other than to save time and cost".

   (c)   Many candidates mentioned a list; database or register and so attained the available mark. However, some did not seem to differentiate between the population and the sampling frame.

   (d)   Most candidates were able to identify the sampling units correctly, although those who had not scored in part (c) tended to say: "the sample of 5 cookers" in part (d).

5.   Almost all candidates answered part (a) correctly, a minority failed to mention "census" or "asking all members" when answers referred to long time/expensive/difficult. In part (b) many candidates failed to include the word "all" in their answer. Quite a number did not know or understand the term sampling frame and wrote about sampling methods. Most candidates answered part (c) correctly, but there were occasional references to golfers rather than members or to those selected in the sample.

6.   In part (a) many candidates were able to calculate the mean accurately, although some divided by random constants. Few drew up a table and many were unable to cope with the 5p coins. The most common error in calculating the variance was the failure to subtract $E(X)^2$.
Most candidates correctly identified 6 possible samples but some failed to realise that combinations such as (1,5) and (5,1) were different and so missed the other 3 possibilities. Only a minority of candidates were able to attempt part (c) of the question with any success, with many candidates having no idea what was meant by 'the sampling distribution of the mean value of the samples'. Most did not find the mean values and if they did, then they were unable to find the probabilities (ninths were common).Very few candidates achieved full marks.

7.   This question proved difficult to many candidates. Errors in this part (a) included the use of the word sample rather than population. Many candidates also gave an ambiguous response to part (b), often omitting to mention all sampling units or the whole population. Part (c) was done badly and whilst some candidates scored 1 mark very few achieved both marks. It appeared that many candidates had attempted to memorise the definition, but it came out garbled and confused with other concepts.

8.   The bookwork required to answer this question was not remembered as well as it should have been. Many candidates could not define a population or a sampling frame in detail or know why they might be different. In part (c) many candidates were unable to give in sufficient detail a justified example of the use of a census and a sample.

9.   Weaker students had difficulties with this question with a considerable number scoring 1 or 0 marks. In part (a) good candidates answered this correctly but for many there was confusion between a population and a sample and that the population must be in a list or equivalent. In part (b) those candidates who had learnt the basic definitions were able to answer this successfully.

10.   Only a very few candidates achieved full marks. Most scored 2 or 3 out of the 5 available. Common errors were in part (c) where only a very small number could provide a valid disadvantage and in part (d) not all candidates realised the problem of having an incomplete (or not up-to-date) sampling frame.

11.   This question also allowed candidates to score highly; indeed some otherwise poor papers were redeemed by good marks here. Most marks were lost in the opening parts where it is clear that candidates do not understand well enough the need for a degree of precision in defining terms such as population and sampling frame. Similarly it is a cause for concern that the majority of candidates talk about a census giving *more accurate* answers (even though this was allowed) rather than understanding the real differences between a sample and a census. Part (e) received a very high number of correct answers, and part (f), although less well done, did receive an encouragingly high number of good solutions, with context being well used. The most common mistakes were careless statements of the hypotheses and a decision to find $P(X = 6)$. Part (g) was very well answered with a large number of candidates gaining full marks. Very few candidates used incorrect parameters in the normal approximation, but the most common cause of loss of marks was in an error in the use of either 70.5 or 82.5 even if a correct probability statement had been given earlier.

12.   No Report available for this question.