Pre-U Mathematics (Statistics with Pure Mathematics) Short Course (1347)

Practice Questions for Paper 2 (Statistics)

1 The table below is based on an official report about the number of people attending hospital Accident and Emergency departments on Mondays to Saturdays in 2008 and 2009.

Day	**2008** ***x*** **(thousands)**	**2009** ***y*** **(thousands)**
Monday	2186	2471
Tuesday	2000	2215
Wednesday	1917	2203
Thursday	1916	2158
Friday	1918	2144
Saturday	1899	2145

Source: HESonline.nhs.uk

$\Sigma x = 11\,836$, $\Sigma x^2 = 23\,409\,466$, $\Sigma y = 13\,336$, $\Sigma y^2 = 29\,720\,000$, $\Sigma xy = 26\,375\,032$

It is believed that there is a linear relationship between x and y.

(i) Calculate the values of S_{xx}, S_{yy} and S_{xy}, giving each correct to the nearest integer. [4]

(ii) Calculate the value of r and show that a linear model is appropriate. [2]

(iii) Calculate the equation of the regression line of y on x. [4]

In 2008, the number of people who attended hospital Accident and Emergency departments on a Sunday was 1958 thousand.

(iv) Use your line to predict how many people attended on a Sunday in 2009. [2]

Mark Scheme:

(i)	$S_{xx} = 23409466 - \frac{(11836)^2}{6} = 60983$ $S_{yy} = 29720000 - \frac{(13336)^2}{6} = 78517$ $S_{xy} = 26375032 - \frac{11836 \times 13336}{6} = 67549$	M1 A1 A1 A1	Correct method for S_{xx} or S_{yy} or S_{xy} Accept 60983 to 60984 Accept 78517 to 78518 Accept 67549 to 67550	**4**
(ii)	$r = \frac{675499}{\sqrt{60983 \times 78517}} = 0.976$ r is near 1, so a good fit to an upward sloping line	M1 A1	Calculating r from their S_{xx}, S_{yy} and S_{xy} (numerical working or their r value correct to 3 sf or better) or using original data Drawing a valid conclusion (confirming that a linear fit is appropriate, as stated in question)	**2**
(iii)	$b = \frac{67549}{60983} = 1.108$ $a = \frac{13336}{6} - 1.108 \times \frac{11836}{6}$ $= 37.6$ $y = 1.108x + 37.6$	M1 M1 A1 A1	Calculating b from their S_{xx}, S_{xy} Allow 1.1 or better, or correct calculation seen Calculating a from Σx, Σy and their b Allow 36.5 to 39 Line with coefficients correct to within tolerances above, without wrong working	**4**
(iv)	$x = 1958$ gives $\hat{y} = 2207$ Predict 2207 thousand	M1 A1	Follow through their line 2150 thousand to 2250 thousand from correct working	**2**

Commentary:

(i) The method mark is for evidence of any one of these calculations being correct. Sight of a formula alone would not be enough, but showing the formula with appropriate numerical values substituted in would be fine, or achieving any one of the values (within tolerance) by using summary statistics or by putting the data through a calculator. The tolerances on the values are to accommodate rounding in the final answer. If the final answers are outside these tolerances and there is no evidence of the correct method having been used then the candidate would score no marks.

(ii) The method mark is for using the values found in part (i) to calculate r (provided the values of S_{xx} and S_{yy} are both positive), or for a fresh start, or for getting the value of r from a calculator. Provided this mark has been awarded, the second mark is then for a valid statement to confirm that a linear fit is appropriate. If a linear fit is not appropriate for the given r value then the second mark is not available. This can be very informal and does not require a formal hypothesis test or use of tables. If a candidate does not gain the first mark then they cannot have the second mark either.

(iii) The method marks are for having the correct method for calculating the gradient and intercept of the regression line. As in part (i), sight of the formula alone would not be enough, but showing the formula with appropriate numerical values substituted or achieving values within the given tolerances is fine. If no method is shown then the final answers must be within the given tolerances. The tolerance on the gradient means 1.1 or 1.11 or 1.108 (but not some other value that has then been rounded to give 1.1). The tolerance on the value of the intercept is to allow for minor rounding errors on the earlier values. The follow through only applies to the method marks.

(iv) The follow through only applies to the method mark. A correct line leading to the answer 2207 (without ‘thousand’, in some form) would get M1 only.

2 The table shows the number of people who successfully quit smoking on 'stop smoking' programmes and the cost per quitter, each year from 2003 to 2010.

Year	Number who successfully quit smoking	Rank	Cost per quitter (£)	Rank
2003	204 876	1	177	5
2004	298 124	2	158	1
2005	329 681	4	159	2
2006	319 720	3	160	3
2007	350 800	6	173	4
2008	337 054	5	219	6
2009	373 954	7	224	8
2010	383 548	8	220	7

Source: The Health and Social Care Information Centre.

(i) Find the median and quartiles of the costs per quitter and use these values to show that there are no years for which the cost per quitter is an outlier. [5]

(ii) Use appropriate calculations from part **(i)** to show that the distribution of the cost per quitter seems to have positive skew. [3]

It has been suggested that the cost per quitter increases with the number of quitters.

(iii) Calculate the value of Spearman's rank correlation coefficient between the number who successfully quit smoking and the cost per quitter. [4]

The critical values for Spearman's rank correlation coefficient when $n = 8$ are given below:

1-Tail Test	5%	2.5%	1%	0.5%
2-Tail Test	10%	5%	2%	1%
$n = 8$	0.6429	0.7381	0.8333	0.8810

(iv) Stating the null and alternative hypotheses being tested, interpret the value obtained in part **(iii)**. [3]

Mark Scheme:

(i)	Median = average of 4th and 5th values = £175 Quartiles = £159.50 and £219.50 IQR = £60 159.50 –90 = 69.50, 219.50 + 90=309.50 All values are between £69.50 and £309.50, so there are no outliers	B1 B1 M1 A1 B1	175 cao Accept 159 to 160 and 219 to 220 Their IQR calculated Fences calculated for their quartiles, or equivalent (159.5-158)/60 = 1.5/60 = 0.025, (224-219.5)/60 = 4.5/60 = 0.075 Explaining how calculations show that there are no outliers	**5**
(ii)	159.50 – 158.00 = 1.50 175.00 – 159.50 = 15.50 175.00 – 158.00 = 17.00 219.50 – 175.00 = 44.50 224.00 – 219.50 = 4.50 224.00 – 175.00 = 49.00 The data values above the median are more ‘spread out’ than those below	M1 M1 A1	Any one of these six calculations, using their values for median and quartiles The corresponding calculation for the other tail Explaining how the calculations show that there is positive skew	**3**
(iii)	1 2 4 3 6 5 7 8 or 1 2 3 4 5 6 7 8 5 1 2 3 4 6 8 7 5 1 3 2 6 4 8 7 -4 1 2 0 2 -1 -1 1 -4 1 0 2 -1 2 -1 1 $\sum d^2 = 28$ $r_s = 1 - \frac{6 \times 28}{8 \times 63} = 1 - 0.333$ $= 0.667$ (3 sf)	M1 A1 M1 A1	Substantially correct calculation of d or $\lvert d \rvert$ or d^2 for the ranks Correct calculation of r_s for their $\sum d^2$ Correct value, to 3 sf or better (or fraction $\frac{2}{3}$)	**4**
(iv)	H_0: no association H_1: positive correlation between ranks 0.6429 < 0.667 Reject H_0 at 5% level (or 10% level if 2-sided) The evidence supports the suggestion that the greater the number of people who quit the higher the cost per quitter	B1 M1 A1	Appropriate statement of hypotheses in any form (allow a two-sided alternative) Or one of 0.7381> 0.667, 0.8333> 0.667 or 0.8810 > 0.667 leading to Accept H_0 at 2.5 % level (or better) or equivalent if 2-sided used A valid conclusion in context	**3**

Commentary:

(i) There are various interpretations of how to find the quartiles for a small data set, and any reasonable interpretation will be credited. The interquartile range and fences, or equivalent, will be follow through from reasonable quartiles. Values that are more than 1.5 times the interquartile range beyond the upper and lower quartiles will be regarded as outliers.

(ii) Positive skew occurs when there is a long tail on the right hand side of the distribution. This could be evidenced by showing that the difference between the median and the lower extreme value is much smaller than the difference between the median and the upper extreme value ($Q_2 - Q_0 << Q_4 - Q_2$); or by showing that the difference between the lower quartile and the lower extreme value is much smaller than the difference between the upper quartile and the upper extreme value ($Q_1 - Q_0 << Q_4 - Q_3$); or by showing that the difference between the median and the lower quartile is much smaller than the difference between the median and the upper quartile ($Q_2 - Q_1 << Q_3 - Q_2$).

(iii) In this question the rank orders were given; if the question does not indicate which way the values are to be ranked then either order will be acceptable. There is no follow through on the accuracy marks in this part.

(iv) 'It has been suggested that the cost per quitter increases with the number of quitters' so a one-tailed test is appropriate; however, in this particular instance there is no penalty for using a two-tailed test. The tables are given in the question because they are not given in the formula book, although the syllabus allows for this type of hypothesis test: 'interpret correlation coefficients in the context of hypothesis tests'.

3 In a large city, the proportion of the population that supports the construction of a new sewer is denoted by p. The water board wants to estimate the value of p.

(i) The water board proposes writing to all households and inviting members of each household to express their views by writing back.

(a) Explain why this method is unsatisfactory. [2]

(b) Suggest an improved method, explaining the advantages of your method. [4]

The water board obtains a random sample of n members of the population, and R of these members support the construction of the sewer.

(ii) Explain why a binomial distribution is likely to be a good model for R. [2]

Assume now that a binomial model is valid for R.

(iii) In the case $R \sim \text{B}(20, 0.4)$, use tables to find

(a) $\text{P}(R > 7)$, [2]

(b) $\text{P}(R = 8)$. [2]

(iv) Now consider the case $R \sim \text{B}(20, 0.23)$

(a) Use the formula to find $\text{P}(R < 3)$, showing all necessary working. [3]

(b) Write down the values of $\text{E}(R)$ and $\text{Var}(R)$. [2]

Mark Scheme:

(i)(a)	Biased (in favour of those with strong views, etc) as mostly only those with strong opinions will reply	M1 A1	"Biased" or equivalent stated Justified in context, any sensible reason	**2**
(i)(b)	Obtain list of population and number the people (from 1 to total number) Select using random numbers Ignore numbers outside range Advantages: unbiased, random method allows probabilities to be used, etc	M1 B1 A1 B1	Not just "allocate random numbers" Not just "select numbers randomly" Any one sensible reason	**4**
(ii)	Random sample implies selections independent (population large) and each has equal probability of being selected	B1 B1	Any two of "random sample", "independent selections", "equal probability of being selected" mentioned for one mark each, up to a maximum of two	**2**
(iii)(a)	$1 - P(R \leq 7) = 1 - 0.4159$ $= 0.5841$	M1 A1	Allow $1 - 0.2500$ (= $P(R \geq 7)$) for M1 A0	**2**
(iii)(b)	$P(R \leq 8) - P(X \leq 7) = 0.5956 - 0.4159$ $= 0.1797$	M1 A1	Question asks for use of tables	**2**
(iv)(a)	$0.77^{20} + 20\times0.23\times0.77^{19} +$ $190\times0.23^2\times0.77^{18}$ $= 0.128(437)$	M1 M1 A1	Correct or with one term extra or missing All correct Answer in range 0.128 to 0.129, before rounding	**3**
(iv)(b)	$E(R) = 4.6$ $Var(R) = 3.542$	B1 B1	Correct answer only Allow 3.54	**2**

Commentary:

(i) The main problem with this method is that although the water board are potentially asking every member of the population they are only likely to get replies from households with strong opinions. Acceptable alternatives would be to identify that they are sampling households rather than individuals (wrong sampling units) or that not every member of the population belongs to a household (sampling frame is incomplete).

The improved method needs to be a fairly detailed description of a practical, random sampling method that addresses these problems, including some issues like random numbers that do not correspond to people on the list.

(ii) Candidates need to interpret the conditions for a binomial model in the context of the question.

(iii) In this part candidates are asked to use B(20, 0.4) tables, so there should be some evidence that they have done this. In part (a), sight of $1 - 0.4159$ or 0.5841 would get the first mark. In this question we are also allowing $1 - 0.2500$ or 0.7500 to get the first mark, being $P(R \geq 7)$. For the accuracy mark the answer must be 0.5841 (to 4 decimal places, as given in the tables). In

part (b), sight of P($R \leq 8$) – P($R \leq 7$) or 0.5956 – 0.4159 or 0.1797 would get the first mark, for the accuracy mark the answer must be 0.1797 (given to 4 decimal places) or this value seen and then rounded to give a final answer of 0.180 or 0.1797 seen and the final answer quoted as being in the range 0.1796 to 0.1799.

(iv) In this part candidates are asked to use the formula, and their working should be shown, not just done on a calculator. The results for the mean and variance of a binomial distribution may be quoted from the formula book.

4 A fair 6-sided dice has its faces numbered 1, 2, 3, 4, 4, 4. The dice is thrown twice, and the total of the two scores is denoted by X.

(i) Construct a probability distribution table for X. [3]

(ii) Use your table to show that $E(X) = 6$ and to find the value of $\text{Var}(X)$. [5]

The mean of 50 observations of X is denoted by $\overline{X}$.

(iii) Write down the approximate distribution of $\overline{X}$, stating the value(s) of any parameter(s). [3]
Hence find $P(\overline{X} \geq 6.2)$. [3]

(iv) Explain why, in answering part **(iii)**, the Central Limit Theorem

(a) can be used, [1]

(b) needs to be used. [1]

Mark Scheme:

(i)	x 2 3 4 5 6 7 8 $P(X = x)$ $^1/_{36}$ $^2/_{36}$ $^3/_{36}$ $^8/_{36}$ $^7/_{36}$ $^6/_{36}$ $^9/_{36}$	B1 M1 A1	Recognising that x can take values 2, 3,..., 8 Any four probabilities correct All correct and presented as a probability distribution	**3**
(ii)	$E(X) = \Sigma xP(X = x)$ $= 216/36 = 6$ $E(X^2) = \Sigma x^2P(X = x)$ [= 1392/36] $Var(X) = E(X^2) - \mu^2$ [= 38.67 – 36] $= 8/3$	M1 A1 M1 M1 A1	Correct method for $E(X)$ Allow unsimplified (answer given) Correct method for $E(X^2)$ Subtracting $[E(X)]^2$ Any equivalent form, e.g. 2.67 (correct to at least 3 sf)	**5**
(iii) **(iii)**	N(6, 0.0533) $1 - \Phi\left(\frac{6.19 - 6}{\sqrt{8/150}}\right)$ $= 1 - \Phi(0.8227)$ $= 0.2054$ (or 0.1934 if no cc)	B1 B1 B1 M1 M1 A1	Normal Mean 6 or their $E(X)$ Variance their $Var(X) \div 50$ Standardise using their μ and σ or σ^2 1 - $\Phi(0.8227)$, follow through their μ and σ^2 Need not have continuity correction Allow in range (0.205, 0.206) or (0.193, 0.194)	**3** **3**
(iv)(a)	Sample size is large enough	B1	Large sample	**1**
(iv)(b)	Parent distribution not normal	B1	Distribution of X is not normal	**1**

Commentary:

Statistics questions will usually be set in a context, but the occasional context free question may be set.

(i) Candidates may use a variety of methods to construct the distribution. This could be as simple as adapting the grid method for the sum of the scores on two standard, fair, six-sided dice.

	1	2	3	4	4	4
1	2	3	4	5	5	5
2	3	4	5	6	6	6
3	4	5	6	7	7	7
4	5	6	7	8	8	8
4	5	6	7	8	8	8
4	5	6	7	8	8	8

The probability distribution can then be constructed from this.

(ii) Most of the marks are for using the right method, so candidates should be encouraged to show their working. The question specifically asks candidates to use their table from part (i), so a

candidate who finds the expected score for a single roll and then doubles it to get 6 has not strictly answered the question. However, they would still be credited with the first two marks.

(iii) There is no penalty for not using the continuity correction when applying the central limit theorem with a discrete parent population. If it is used in this case then the value of X is reduced by 0.5 and hence the value of $\bar{X}$ by $0.5 \div 50 = 0.01$.

(iv) The Central Limit Theorem gives the distribution of $\bar{X}$ when the parent population is normally distributed, whatever the sample size. It also gives a good approximation to the distribution of $\bar{X}$ when the parent population is not normally distributed, provided the sample size is large enough. For a large sample the candidates have no easy way of finding the true distribution of $\bar{X}$ and so the Central Limit Theorem is the only method available to them.

5 A shop manager believes that 55% of the shop's customers are female. A test of whether the manager is overestimating the proportion of female customers is carried out at the 5% significance level, based on a random sample of 14 customers.

(i) State appropriate hypotheses for the test, explaining the meaning of any symbol used (other than H_0 and H_1). [3]

(ii) Find the critical region for the test. State the probability of a value being in the critical region, assuming H_0 to be true. [3]

(iii) State the conclusion of the test if 5 of the sample of 14 customers are female. [3]

(iv) If the true proportion of females in the population is 40%, find the probability that the test results in a Type II error. [3]

Mark Scheme:

(i)	$H_0 : p = 0.55$; $H_1 : p < 0.55$ where p is the proportion of all the shop's customers who are female	B1 B1 B1	Null hypothesis correct Alternative hypothesis correct Valid interpretation of p as population proportion	**3**
(ii)	X~ B(14, 0.55) where X is the number of female customers in the sample Critical region is $X \leq 4$ $P(X \leq 4) = 0.0426$	M1 A1 A1	B(14, 0.55) stated or implied This CR stated, or equivalent, with inequality This probability seen	**3**
(iii)	$5 > 4$ so 5 is not in the critical region, or (assuming H_0) $P(X \leq 5) = 0.1187 > 0.05$ Do not reject H_0 (accept H_1) Insufficient evidence to support claim that the proportion of female customers is being overestimated	M1 M1 A1	Explicit comparison of 5 with CR Correct first conclusion, FT their CR from (ii) Conclusion contextualised, including an acknowledgement of uncertainty (e.g. insufficient evidence)	**3**
(iv)	True distribution is now $X \sim B(14, 0.40)$ In this distribution $P(X > 4)$ $= 0.7207$	M1 M1 A1	B(14, 0.4) used or implied P(not in their CR), FT their CR Anything rounding to 0.721	**3**

Commentary:

(i) The null and alternative hypotheses should be stated in terms of a probability p and the definition of p should explicitly refer to the population.

(ii) The critical region should be given as an inequality, or a listing of values.

(iii) The observed value, 5, is not in the critical region and hence we do not have enough evidence to reject H_0. The conclusion should be interpreted in context.

(iv) A Type II error occurs when we accept H_0 although it is actually false. In this case this means that we observe a value of 5 or above in B(14, 0.40).
A Type I error occurs when we reject H_0 although it is actually true. The probability of making a Type I error was calculated in part (ii). There is always a 'trade off' between keeping the probability of a Type I error small and not letting the probability of a Type II error get too large. We could reduce the probability of a Type II error by increasing the sample size.

6 Roz has carried out a study to investigate whether eating breakfast improves school performance. She chose 40 students at random to take part in the study and set them a Maths test on Tuesday morning. She then randomly assigned 20 of the students to each of two groups, B and A. The students in group B were asked to eat breakfast before coming to school the next day, those in group A were asked to avoid breakfast. The twenty students were given a second Maths test on Wednesday morning and the scores on the two tests were compared.

If the scores on the two tests were the same or only differed by one mark, Roz recorded that the scores were unchanged, otherwise she recorded that there had been a decrease or an increase in the score. The number of students in each category is shown below.

		Change in score		
		Decrease	Unchanged	Increase
Group	B	0	2	18
	A	6	8	6

Roz tested for independence in the contingency table using a chi-squared test.

(i) Assuming independence, calculate the expected frequency for the cell corresponding to students in group B whose scores decreased. Explain how the table must be reduced before the test statistic can be calculated. [3]

(ii) Write out the table of observed frequencies, the table of expected frequencies and the corresponding table of chi-squared entries. Hence find the value of the chi-squared statistic for the test. [5]

(iii) Complete the test at the 5% level of significance and state the conclusion. [3]

(iv) Give two criticisms of the design of the experiment. [2]

Mark Scheme:

(i)	Row total = 20, column total = 6 Assuming independence, expected frequency = $40\times\frac{6}{40}\times\frac{20}{40}=3$ Expected frequency < 5 so first two columns must be combined	M1 A1 B1	Row and column totals seen or implied 3 calculated from correct method Value is small so columns must be merged	**3**
(ii)	obs - /0 + exp - / 0 + B 2 18 B 8 12 A 14 6 A 8 12 Yates' correction is required χ^2 - /0 + B 3.781 2.521 $\chi^2 = 12.6$ A 3.781 2.521	B1 M1 A1 M1 A1	Observed frequencies correct (first two columns merged) Expected frequencies correct for their observed frequencies Expected frequencies correct χ^2 calculations correct for their obs and exp (with or without Yates' correction) $\chi^2 = 12.6$ or better, from correct working and Yates' correction used	**5**
(iii)	$\nu = 1$ and $p = 5\% \Rightarrow$ cv = 7.8794 Reject H_0 Evidence supports claim that eating breakfast improves test scores	M1 A1 B1	7.8794 (follow through their table) Correct conclusion for their 12.6 An appropriate conclusion, in words and in context	**3**
(iv)	Small sample size Only tests performance in Maths tests May improve anyway because of having done first test Does not specify what sort of breakfast Does not ask whether they normally eat breakfast	B1 B1	A valid criticism Any other valid criticism	**2**

Commentary:

(i) Alternatively, the expected frequency may be found as 20×6÷40, or equivalent.
The expected frequencies need to be at least about 5, so we must merge cells. Clearly it makes no sense to merge rows, and if we merge columns then they should be adjacent, so we merge the first two columns. The smallest expected frequency is then 8.

(ii) The question specifically asks for the table of observed frequencies and the table of expected frequencies. If a candidate has not merged the first two columns then they cannot get the first mark. Because the reduced table is 2×2, Yates' correction is required, this means that we subtract 0.5 from |O-E| before squaring and dividing by E for each cell. If Yates' correction is not used the last mark is not given, however this should not affect the conclusion in part (iii).

(iii) The reduced table has only 1 degree of freedom and the calculated chi-squared value is considerably greater than the critical value meaning that H_0 is rejected and we have evidence that there is some association between the change in score and whether or not breakfast was eaten. Referring back to the tables of observed and expected frequencies then enables us to draw a conclusion in context.

(iv) The first mark is for recognising any valid criticism of the design, such as that the sample size is small. The second mark is for any other valid criticism of the experimental design, such as one of those given in the mark scheme.

7 The breaking strength of any rope is the maximum load it can hold before it breaks. A manufacturer of nylon ropes knows that the breaking strength of any type of nylon rope may be assumed to follow a normal distribution.

The manufacturer advertises that, for a certain type of nylon rope, 90% of the ropes can hold a load of 1200 Newtons and 99.9% of the ropes can hold a load of 1087 Newtons.

(i) Calculate the values of the population mean and standard deviation that are consistent with this information. [5]

Ten nylon ropes of a different type are tested and their breaking strengths, in Newtons, are found to be:

993 1000 1012 1021 1034 1038 1043 1047 1054 1058

(ii) Calculate a 95% confidence interval for the population mean for the distribution of breaking strengths for this type of rope. [8]

Mark Scheme:

(i)	$-1.282 = \frac{1200-\mu}{\sigma}$ and $-3.090 = \frac{1037-\mu}{\sigma}$ $\Rightarrow 1200-\mu = -1.282\sigma$ and $1087-\mu = -3.090\sigma$ $\Rightarrow \sigma = 62.5$ N and $\mu = 1280.125$ N	B1 B1 M1 A1 A1	–1.282 –3.090 *z* values used appropriately $\sigma = 62.5$ $\mu = 1280.125$ (given to at least 3 sf)	**5**
(ii)	$\Sigma x = 10300 \Rightarrow \bar{x} = 1030$ $\Sigma x^2 = 10613572 \Rightarrow \Sigma(x-\bar{x})^2 = 4572$ $\Rightarrow$ unbiased estimator for $\sigma^2 = \frac{4572}{9} = 508$ Small sample and estimated variance *so* $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \sim t(n-1)$ Critical values in t(9) are $\pm$ 2.262 $\frac{1030-\mu}{\sqrt{\frac{308}{10}}} = \pm 2.262$ $\Rightarrow$ CI for $\mu = (1013.9, 1046.1)$	B1 M1 A1 A1 M1 A1 M1 A1	1030 Correct method used (or from calculator) 5508 612 *t* distribution with 9 degrees of freedom 2.262 Correct method used Confidence limits correct to 3sf or better	**8**

Commentary:

(i) This is the kind of information that is given in everyday situations, such as rope strengths. By using a normal model with unknown mean and standard deviation, and using the reverse normal probability tables, we can obtain simultaneous equations in μ and σ and then solve these to find values for the parameters. The values should be given to at least 3 significant figures.

(ii) This question concerns the distribution of the sample mean for a sample drawn from a normal population. As the sample is small and the population variance has been estimated, we should use a *t* distribution to model the standardised variable.

8 It has been suggested that the median household in Britain 1911 consisted of 3.5 people. A researcher thinks that this figure is too low. A summary of an extract from 20 households, chosen at random from the 1911 census, is given below.

Household number	Number of males	Number of females	Number of persons
1	3	1	4
2	3	3	6
3	4	1	5
4	2	4	6
5	8	3	11
6	3	6	9
7	1	1	2
8	3	2	5
9	2	1	3
10	3	1	4
11	6	2	8
12	2	2	4
13	5	3	8
14	2	2	4
15	3	4	7
16	1	1	2
17	1	3	4
18	5	5	10
19	0	1	1
20	4	1	5

(i) Use a sign test, at the 10% level of significance, to test the claim against the alternative that the median household size was greater than 3.5. [6]

(ii) Give two reasons why it would be inappropriate to use a Wilcoxon signed rank test on this data. [2]

Mark Scheme:

(i)	H_0: $p = 0.5$ p = P(household size > 3.5) H_1: $p > 0.5$ one-sided test X = number of households of > 3.5 people Assuming H_0: $X \sim B(20, 0.5)$ In B(20, 0.5), P($X \geq 16$) = 1 – 0.9941 = 0.0059 0.0059 < 0.05 Reject H_0, the data support the claim that the median household size is greater than 3.5	B1 B1 M1 A1 M1 A1	An appropriate definition of p Null and alternative hypotheses correct B(20, 0.5), seen or implied In B(20, 0.5), critical value (1-sided 5%) is 14 16 > 14 Valid conclusion, in context, from correct working and one-sided alternative used	**6**
(ii)	Because of tied ranks Would need to be able to assume a symmetric population	B1 B1	Several of the values are the same A valid statement about the shape of the population distribution	**2**

Commentary:

(i) Sometimes tables of data will be given that may involve unnecessary information. Extracting the required information is part of the task for the person analysing the data. Here the required data is in the final column – household size.

(ii) Although a Wilcoxon test may be carried out with tied data, it makes for a lot more work and is excluded from the specification other than in the very simplest cases. Wilcoxon tests are non-parametric tests assuming only that the underlying population is symmetric.

9 A supermarket sells cheese whose masses have the distribution N(μ, σ^2). The following table gives the masses of a random sample of 16 cheeses.

989	990	991	992	993	994	995	996
997	998	999	1000	1001	1002	1003	1004

(i) Obtain unbiased estimates of μ and σ^2. [4]

(ii) Test at the 1% significance level whether the population mean mass of the cheese is 1000. [7]

(iii) Determine whether the conclusion of the test would have been different if the value of σ^2 was known to be $\frac{68}{3}$. [3]

Mark Scheme:

(i)	$\hat{\mu} = \bar{x} = 996.5$ $\Sigma x^2/16 - 996.5^2$ [= 21.25] $\times$ 16/15 = 22.67	B1 M1 M1 A1	Mean 996.5 Correct method for variance Biased estimate, multiply by 16/15 Anything that rounds to 22.7	**4**
(ii)	$H_0 : \mu = 1000$ $H_1 : \mu \neq 1000$ where μ is the population mean mass of cheese $t_{15} = \frac{996.5 - 1000}{\sqrt{22.6667/16}} = -2.94$ ≤ -2.947 Do not reject H_0 Insufficient evidence that mean mass of cheese is not 1000	B1 B1 M1 A1 B1 M1 A1	Null hypothesis correct, using μ Two-tailed alternative Use of X, $\bar{x}$, $\bar{X}$ or 996.5 gets B0 B0. Standardise with 16, allow $\sqrt{}$ errors Correct t (must be –ve) Explicit comparison with –2.947 or –2.95 or equivalent Correct conclusion from μ and $\bar{x}$ the right way round and correct use of $\sqrt{n}$ Conclusion in context, must acknowledge uncertainty (e.g. mention "evidence")	**7**
(iii)	Can now use CLT –2.94 < –2.576 So we now reject H_0 and make a different conclusion from before (mean less than 1000)	M1 A1 A1	Use normal not t Correct comparison, or equivalent Correct conclusion, no need for context/ evidence here	**3**

Commentary:

(i) There are various ways to achieve the unbiased estimate for σ^2.

(ii) The parameter, μ, should be defined when setting up a null hypothesis. Here μ has been defined in the question (provided the same symbol is used). The alternative hypothesis is two tailed because the question asks whether the mean is 1000 (or not). If a candidate uses the sample mean or uses the value of the sample mean then they lose these two marks.

The t distribution is used because we are testing the value of a mean with a small sample from a normal distribution and the variance has been estimated.

Equivalently, the critical values may be calculated as $1000 \pm 2.947\sqrt{(22.67/16)} = 996.49$ and 1003.51, 996.5 falls between these and hence the conclusion. Note the need for accuracy as the value is very close to being critical.

(iii) The variance is now known, so we can use the Central Limit Theorem.
Equivalently, $1 - \Phi(2.94) = 0.0015 < 0.005$ and hence reject H_0 (or accept H_1).

10 **(i)** Explain when a non-parametric significance test for the value of a sample median should be used. [2]

A farmer believes that organically-grown apples have larger median mass than non-organically-grown apples. The masses of a random sample of 7 organically-grown apples and 9 non-organically-grown apples were obtained. The rankings of these masses were found to be as follows (1 = largest mass, 16 = smallest mass):

Organically grown:	1	2	5	6	10	11	14		
Non-organically-grown:	3	4	7	8	9	12	13	15	16

(ii) Use an appropriate non-parametric test at the 5% significance level to test the farmer's belief. [7]

(iii) A further three non-organically-grown apples are included in the sample. All are found to have masses less than the smallest mass in the original sample of 16 apples. Carry out a revised test based on the masses of all 19 apples, using a normal approximation. [7]

Mark Scheme:

(i)	When the parent distribution is not known to be normal and when the sample size is not large enough for the Central Limit Theorem to be used	B1 B1	Parent distribution not known Small sample	**2**
(ii)	$H_0 : M_o = M_n$ $H_1 : M_o > M_n$ where $M_o = M_n$ are the population median masses of organically-grown and non-organically-grown apples respectively $R_m = 49$ $m = 7, n = 9$ $m(m + n + 1) – R_m = 70$ $\Rightarrow W = 49$ CV 43 (from tables) Do not reject H_0. Insufficient evidence that organically-grown apples have higher mass	B1 B1 M1 A1 B1 M1 A1	Must use population medians Inequality between population medians correct (may be described using a difference of the population medians) Substantially correct calculation of R_m $R_m = 49$ and $m(m + n + 1) – R_m$ considered Critical value 43 stated Correct conclusion In context, acknowledge uncertainty (e.g. mention “evidence”)	**7**
(iii)	R_m still 49 N(70, 140) $Z = (49.5 – 70)/\sqrt{140})$ $= –1.732$ $< – 1.645$ Reject H_0, etc	M1 B2 M1 A1 B1 M1	R_m unchanged Correct parameters for normal Standardise, allow no cc, allow √ error Correct including continuity correction Explicitly compare with –1.645 (or p with 0.05) Correct conclusion (no need for context/ evidence here)	**7**

Commentary:

(i) A non-parametric test would not be needed if the parent population were known to be normal (since we could test an average using the population mean with a t distribution if the variance had been estimated or CLT if the variance was known), and it would not be needed if the sample size were large (since we could test the population mean using CLT).

Ideally the parent population should be fairly symmetric.

(ii) The important thing here is to show evidence of calculating W as the smaller of R_m and $m(m + n + 1) – R_m$.

(iii) This tests the use of the normal approximation $W \sim \mathrm{N}\left(\frac{1}{2}m(m+n+1), \frac{1}{12}mn(m+n+1)\right)$. Since W is discrete, a continuity correction is required.

Copyright Acknowledgements:

Question 1 © Provisional Accident and Emergency Quality Indicators for England; October 2011; http://ww.ic.nhs.uk/statistics-and-data-collections/hospital-care/accident-and-emergency-hospital-episode-statistics-hes/provisional-accident-and-emergency-quality-indicators-for-england-experimental-statistics-by-provider-for-may-2011.
© Statistics on NHS Stop Smoking Services: England, April 2010 to December 2010.

Question 2 April 2011; http://www.ic.nhs.uk/statistics-and-data-collections/health-and-lifestyles/nhs-stop-smoking-services:-england-april-2007-to-march-2008-annual-report.

Question 2 © Statistics on NHS Stop Smoking Services: England, April 2007 to March 2008. August 2008; http://www.ic.nhs.uk/statistics-and-data-collections/health-and-lifestyles/nhs-stop-smoking-services/statistics-on-nhs-stop-smoking-services:-england-april-2007-to-march-2008-annual-report.

University of Cambridge International Examinations is part of the Cambridge Assessment Group. Cambridge Assessment is the brand name of University of Cambridge Local Examinations Syndicate (UCLES), which is itself a department of the University of Cambridge.