Question 1

A random sample of 18 software engineers was selected and it was found that their average income was \$40,000 with a standard deviation of \$3,125.

Calculate the following:

(a) An 80% confidence interval for μ. **(b)** A 95% confidence interval for μ. **(c)** A 99% confidence interval for μ.

Question 2

Guinness brewery wish to compare the quality of stout made using two different varieties of barley. Samples of the drink were prepared and subsequently tested. Taking various factors into consideration, each one was then given an overall quality score (where a higher score indicates better quality). The results are as follows:

Variety1	10	8	7	8	6	
Variety2	5	6	8	6	7	7

The aim is to compare the mean scores in the two groups.

(a) If we wish to make the equal variance assumption in our calculation - what test must we carry out? **(b)** By carrying out this test, show that the equal variance assumption is reasonable here. **(c)** Calculate a 95% confidence interval for the difference between the two means (using the equal variance approach). State your conclusion. **(d)** What is the advantage of the *unequal* variance approach? Calculate a 95% confidence interval using this approach.

Question 3

Seven athletes were asked to run 100m without warming up prior to running. On another day they warmed up first and then ran. On both occasions they were timed and the results (in seconds) are as follows:

Individual	1	2	3	4	5	6	7
No Warm Up	13.6	12.8	12.3	11.7	12.0	13.3	10.5
Warm Up	13.9	12.4	12.2	11.6	11.9	12.7	10.4

(a) Calculate a 95% confidence interval for the *average difference* in times and hence comment on the usefulness of warming up (hint: the data is paired).

1. The mean height of the women in a large population is 1.671m while the mean height of the men in the population is 1.758m. The mean height of all the members of the population is 1.712m. Calculate the percentage of the population who are women.

2. An analyst of the retail trade uses as analytical tools the concepts of Footfall (the daily number of customers per unit sales area of a shop) and Ticket Price (the average sale price of an item in the shops offer).
   Shops are classified as offering Low, Medium or High price items and, during any sales period, as having Low, Medium or High footfall.
   During the January Sales the analyst studies a sample of shops and obtains the following frequency data for the nine possible combined classifications:

	Low Price	Medium Price	High Price
Low Footfall	45	75	23
Medium Footfall	37	126	25
High Footfall	22	43	16

   Conduct a suitable test for association between Ticket classification and Footfall level, and report on your findings.

Inference

1. In the past, 18% of shoppers have bought a particular brand of breakfast cereal. After an advertising campaign, a random sample of 220 shoppers is taken and 55 of the sample have bought this brand of cereal.

   - Write down the null and the alternative hypothesis for this problem
   - State whether it is a one tailed or two tailed test

2. The starting annual salaries for students graduating from two departments, X and Y, of a university are being investigated. Two random samples of last years intake have been selected and the results are as follows:

Dept	Sample size	Mean starting salary	Std. Deviation
X	50	21,000	4,900
Y	40	20,000	3,000

   - What proportion of new graduates from Department X earn more than 22,000 per month?
   - Test the hypothesis that graduates from Department X earn more than those from Department Y at two appropriate levels and comment on your results. Give any necessary conditions for your test to be valid.

3. A market research company has conducted a survey of adults in two large towns, either side of an international border, in order to judge attitudes towards a controversial internationally broadcast celebrity television programme. The following table shows some of the information obtained by the survey:

	Town A	Town Z
Sample size	50	50
Sample number approving	26	22

   - Conduct a formal hypothesis test, at the 5% significance level, of the claim that the population proportions approving the programme in the two towns are equal.
   - Would your conclusion be the same if, in both towns the sample sizes had been 100 (with the same sample proportions of approvals)?

4. The intelligence quotient (IQ) of 36 randomly chosen students was measured. Their average IQ was 109.9 with a variance of 324. The average IQ of the population as a whole is 100.

   - Calculate the p-value for the test of the hypothesis that on average students are as intelligent as the population as a whole against the alternative that on average students are more intelligent.
   - Can we conclude at a significance level of 1% that students are on average more intelligent than the population as a whole?
   - Calculate a 95% confidence interval for the mean IQ of all students.

5. The manufacturer of the new spray also claims that it can be used to prevent the loss due to insect damage of tender seedlings. To test this claim, the grower sprays 50 tomato seedlings with the new spray and his remaining 100 tomato seedlings with his standard spray. After six weeks, the fruit grower counts the number of healthy plants with the following results.

	New spray	Standard spray
No. of seedlings sprayed	50	100
No. of healthy plants at 6 weeks	40	70

Construct an approximate 95% confidence interval for the difference in the proportion of healthy plants six weeks after spraying between the two groups.

Question 1 (Paired t-test)

The weight of 10 students was observed before commencement of their studies and after graduation (in kgs). By calculating the realisation of the appropriate test statistic, test the hypothesis that the mean weight of students increases during their studies at a significance level of 5%.

Student	1	2	3	4	5	6	7	8	9	10
Weight before	68	74	59	65	82	67	57	90	74	77
Weight after	71	73	61	67	85	66	61	89	77	83

[**Recall Descriptive Statistics**]
You may be required to carry out these calculations in the exam.

- Case-wise differences are
$$d = \{3, -1, 2, 2, 3, -1, 4, -1, 3, 6\}$$
- The sum of case-wises differences and squared case-wise differences are $\sum d_i = 20$ and $\sum d_i^2 = 90$ respectively.
- Mean of case-wise differences $\bar{d} = 2.00$.
$$\bar{d} = \frac{3 + (-1) + 2 + \ldots + 6}{10} = \frac{20}{10}$$
- Standard deviation of casewise differences $s_d = 2.36$
(Modified version of standard deviation formula)
$$s_d = \sqrt{\frac{\sum(d_i^2) - \frac{(\sum d_i)^2}{n}}{n-1}}$$
$$s_d = \sqrt{\frac{90 - \frac{(20)^2}{10}}{9}} = \sqrt{\frac{50}{9}} = 2.36$$
- Standard Error
$$S.E.(d) = \frac{s_d}{\sqrt{n}} = \frac{2.36}{\sqrt{10}} = 0.745$$
- From Murdoch Barnes, the CV is 1.812 (small sample,df = 9,one-tailed procedure)

Writing the Hypotheses

H_0 $\mu_d \leq 0$
mean of case-wise differences not a positive number. (i.e. no increase in weight)

H_1 $\mu_d > 0$
mean of case-wise differences is a positive number. (i.e. increase in weight)

Question 1 Part B

Calculate a 95% confidence interval for the amount of weight that students put on during their studies. Using this confidence interval, test the hypotheses that on average students put on **3 kilos** during their studies

Question 2 (Two Sample Means - One Tailed)

A pharmaceutical company wants to test, a new medication for blood pressure. Tests for such products often include a '*treatment group*' of people who use the drug and a '*control group*'of people who did not use the drug. 50 people with high blood pressure are given the new drug and 100 others, also with high blood pressure, are not given the drug.

The systolic blood pressure is measured for each subject, and the sample statistics are given below. Using a 0.05 level of significance, test the claim that the new drug **reduces** blood pressure.

Treatment	Control
$n_1 = 50$	$n_2 = 100$
$x_1 = 189.4$	$x_1 = 203.4$
$s_1 = 39.0$	$s_1 = 39.4$

Standard Error Formula

$$S.E.(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Question 3 (Two Sample Means - One Tailed)

The average mass of a sample of 64 Irish teenagers (Let say - 18 year old males) was 73.5kg with a variance of 100kg^2. The average mass of an equivalent sample of 81 Japanese teenagers was 68.5kg with a variance of $81\ \text{kg}^2$.

(i) Test the hypothesis that Irish students are larger (in terms of mass) than Japanese teenagers.

Standard Error Formula

$$S.E.(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$S.E.(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{10^2}{64} + \frac{9^2}{81}} = \sqrt{2.56} = 1.6$$

Question 4 (Two Sample Means, small samples, one tailed)

The working lifetimes of 100 of both of two different types of batteries were observed. The mean lifetime for the sample of type 1 batteries was 25 hrs with a standard deviation of 4hrs. The mean lifetime for the sample of type 1 batteries was 23 hrs with a standard deviation of 3hrs.

Type 1	Type 2
$n_1 = 100$	$n_2 = 100$
$x_1 = 25$ hours	$x_1 = 23$ hours
$s_1 = 4$ hours	$s_1 = 3$ hours

(i) Test the hypothesis that the mean working lifetimes of these batteries do not differ at a significance level of 5% .

(ii) Calculate a 95% confidence interval for the difference between the average working lifetimes of these batteries.

(iii) Using this confidence interval, test the hypothesis that battery 1 on average works for 3 hours longer than battery 2.

Question 5 (Two Sample proportions, one tailed)

A simple random sample of front-seat occupants involved in car crashes were taken. The first sample was on cars with airbags available and it was found that there were 29 occupant fatalities out of a total of 1110 occupants. The second sample was on cars with no airbags available and there were 62 fatalities out of a total 1553 occupants.

(i) Using a 5% significance level, determine whether or not there is a difference in the proportion of fatality rates of occupants in cars with airbags and cars without airbags.

(ii) Calculate a 95% confidence interval for the difference between the two proportions of fatality rates.

Standard Error Formula
Confidence Intervals

$$S.E.(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 \times (100 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (100 - \hat{p}_2)}{n_2}}$$

Hypothesis testing

$$S.E.(\pi_1 - \pi_2) = \sqrt{\bar{p} \times (100 - \bar{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Aggregate Sample Proportion

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Confidence Intervals (in terms of percentages)
95% confidence interval

$$(\hat{p}_1 - \hat{p}_2) \times (1.96 \times S.E.(\hat{p}_1 - \hat{p}_2))$$

$$1.4 \times (1.96 \times 0.683) = (1.27, 1.53)$$

Question 6 (Two Sample proportions, one tailed)

- The government wishes to increase the proportion of people taking government training courses who obtain a job in the following 3 months.
- Before they introduced the new schemes this figure was 58%, according to sample of 400 people, with 232 successes.
- A survey of 300 people who took the new courses indicated that 188 of them gained a job. A government official stated that this indicates that the new courses have been more successful.
- Is this statement reasonable at a significance level of 5%?

Some Calculations

- Aggregate proportion

$$\bar{p} = \frac{232 + 188}{400 + 300} = \frac{420}{700} = 60\%$$

- Standard Error for Hypothesis Test

$$S.E$$

$$S.E.(\pi_1 - \pi_2) = \sqrt{60 \times 40) \times \left(\frac{1}{400} + \frac{1}{300}\right)} = 3.74$$

Question 7 - Two Sample Means (Small Samples)

A new process has been developed to reduce the level of corrosion of car bodies.

- Experiments were carried out on 11 cars using the new process and 11 cars using the old process.
- The average level of corrosion using the new process was 3.4 with a standard deviation of 0.5.
- The average level of corrosion using the old process was 4.2 with a standard deviation of 0.8.

(i) Test the hypothesis that the variance of the level of corrosion does not depend on the process used.

(ii) Is there any evidence that the new process is better at a significance level of 5%?

(iii) Calculate a 95% confidence interval for the difference between the mean levels of corrosion under the two processes. Can it be stated that the mean level of corrosion is reduced by 1.5 at a significance level of 5%?

Question 8 - Two Sample Means

Deltatech software has 350 programmers divided into two groups with 200 in Group A and 150 in Group B. In order to compare the efficiencies of the two groups, the programmers are observed for 1 day.

- The 200 programmers in Group A averaged 45.2 lines of code with a standard deviation of 8.4.
- The 150 programmers in Group B averaged 42.7 lines of code with a standard deviation of 5.2.

Let $\bar{x}_A$ denote the average number of lines of code per day produced by programmers in Group A and let $\bar{x}_A$ be the corresponding statistic for Group B. Provide an estimate of $\mu_A?\mu_B$ and calculate an approximate 95% confidence interval for

Test the claim that Group A are more efficient than Group B by

(i) Interpreting the 95% confidence interval.

(ii) Computing the appropriate test statistic.

Question 9 - Two Sample Proportions

In a recent British election 40.12% of the voters voted for the Labour party. A survey of 98 people indicated that 49 of them wish to vote for the Labour party.

(i) Does this figure indicate that support for the Labour party has changed at a significance level of 5% (calculate the realisation of the appropriate test statistic)?

(ii) Calculate a 95% confidence interval for the present support of the Labour party. Comment on your result taking your conclusion from part i) into account.

Question 10 - Testing Equality of Variances

Interpret the output from the following tests of equality of variances. State your conclusion both by referencing the $p-$value and the confidence interval. You may assume the significance level is 5%.

(Remark : This procedure is a one-tailed procedure. However, we will base our conclusion on whether or not we arbitrarily decide the p-value is large or small)

```
> var.test(X,Y)

F test to compare two variances

data:  X and Y
F = ???, num df = 13, denom df = 13, p-value = 0.02725
alternative hypothesis:
true ratio of variances is not equal to 1
95 percent confidence interval:
1.164437 11.299050
sample estimates:
ratio of variances
????
```

```
> var.test(X,Z)

F test to compare two variances

data:  X and Z
F = ???, num df = 13, denom df = 11, p-value = 0.7813
alternative hypothesis:
true ratio of variances is not equal to 1
95 percent confidence interval:
0.2526643 2.7401535
sample estimates:
ratio of variances
??????
```

Question 11 - Shapiro-Wilk Test

Interpret the output from the three Shapiro-Wilk tests. What is the null and alternative hypotheses? State your conclusion for each of the three tests.

```
> shapiro.test(X)

Shapiro-Wilk normality test
```

```
data:  X
W = 0.9001, p-value = 0.113
>
```

```
> shapiro.test(Y)

Shapiro-Wilk normality test

data:  Y
W = 0.8073, p-value = 0.006145
>
```

```
> shapiro.test(Z)

Shapiro-Wilk normality test

data:  Z
W = 0.9292, p-value = 0.372
```

Question 12 - Classification Metrics

For each of the following classification tables, calculate the following appraisal metrics.

- accuracy
- recall
- precision
- F-measure

	Predict Negative	Predict Positive
Observed Negative	9500	85
Observed Positive	115	300

	Predict Negative	Predict Positive
Observed Negative	9700	140
Observed Positive	60	100

	Predict Negative	Predict Positive
Observed Negative	9530	10
Observed Positive	300	160

Question 6

In a study of company salaries, salaries paid by 2 different IT companies were randomly selected.

- For 40 Deltatech employees the mean is 23,870 and the standard deviation is 2,960.

- For 35 Echelon employees , the mean is 22,025 and the standard deviation is 3,065.

At the 0.05 level of significance, test the claim that Deltatech employees earn the same as their Echelon counterparts.

Question 7

Does it pay to take preparatory courses for standardised tests such as the Comptia Exams?

Using the sample data in the following table, compute the case-wise differences, the mean of the case-wise differences and the standard deviation of the case wise differences for the following data set.

Student	A	B	C	D	E	F	G	H	I	J
Score Before	700	840	830	860	840	690	830	1180	930	1070
Score After	720	840	820	900	870	700	800	1200	950	1080

Question 1

A machine is set to produce laptop sleeves with the following dimensions:
length $\times$ width $\times$ depth = 40 cm $\times$ 30 cm $\times$ 2 cm. A sample of 40 sleeves was selected and each was measured. The results were as follows:

	length	width	depth
$\bar{x}$	40.11	30.09	1.91
s	0.51	0.17	0.15

Test the following hypotheses (use the 5% level of significance in each case):

(a) The mean length is equal to 40cm. **(b)** The mean width is equal to 30cm.
(c) The mean depth is equal to 2cm. **(d)** Both the width and the depth of the sleeve need to be addressed here - which do you think is more urgent? **(d)** Calculate the p-values for the tests carried out in parts (a), (b) and (c).

Question 2

A matchbox is supposed to contain 100 matches. We wish to test this hypothesis.

(a) State the null and alternative hypotheses. **(b)** From a sample of 32 matchboxes, it is found that the average is 99.4 and the standard deviation is 2.1. Calculate the test statistic. **(c)** Provide your conclusion based on the p-value.

Question 3

An aircraft part is designed to last more than 500 hours. However, in the interest of safety, it will first be assumed that the part lasts *less than or equal to* 500 hours (i.e., this is the null hypothesis) unless there is firm evidence suggesting otherwise.

(a) State the null and alternative hypotheses. **(b)** What is the critical value if $\alpha = 0.001$ and only 4 units will be run until wearout (due to the expense of wasting aircraft parts).
(c) In this sample of size 4, it is found that the average is 566 hours and the variance is 83 hours^2. Calculate the test statistic. **(d)** What is the conclusion?

Question 4

A friend claims that he can pass a particular game in 4 hours or less (on average). We wish to test this hypothesis at the 10% level of significance. Your friend plays the game on 6 different occasions: his average completion time is 4.6 hours and the standard deviation is 0.5 hours.

(a) State the null and alternative hypotheses. **(b)** What is the critical value? **(c)** Calculate the test statistic and provide your conclusion. **(d)** Between what two values does the p-value lie? (note: the p-value cannot be calculated exactly using the t-tables)

Question 5

A die is rolled 80 times and we count 18 sixes. We wish to test the hypothesis that the die is fair (note: if this is the case, the proportion of sixes is $p = \frac{1}{6}$).

(a) State the null and alternative hypotheses. **(b)** If we wish to test at the 5% level of significance, what is the critical value? **(c)** Calculate the test statistic and provide your conclusion.

Question 6

Assume that a particular brand dominates the market. More specifically, it is well-known that at least 60% of people use this brand (i.e., $p \geq 0.6$). However, in response to recent media claims that this brand is weakening, the company wish to test the hypothesis that $p \geq 0.6$.

(a) State the null and alternative hypotheses. **(b)** From a sample of 1000 people, it is found that 629 use this brand; calculate the test statistic and, hence, the p-value. **(c)** Based on the evidence, state your conclusion.

Question 7

Last year 30% of applicants to a graduate programme failed the aptitude test. This year 100 graduates applied - 25% of these failed the test.

(a) We wish to test the hypothesis that the quality of applicants has not changed since last year - what are the null and alternative hypotheses? **(b)** If we are testing at the 1% level, what is the rejection region? **(c)** Based on the data, calculate the test statistic and provide your conclusion.

Question 1

A market researcher wishes to know the market share for Android devices. From a sample of 500 individuals, it was found that 359 use an Android device.

(a) What type of data has been collected here? **(b)** What is the parameter and its value? **(c)** What is the statistic and its value? **(d)** Calculate a 95% confidence interval and interpret this interval. **(e)** How large a sample is required to reduce the *margin of error* in the previous confidence interval to ± 0.02?

Question 2

Google want to estimate the average amount (in dollars) that an individual spends after clicking on a particular Google Ad. In order to achieve this, 1000 people are randomly selected and the amount they spend is recorded. It is found that the average spend is \$42.38 and the standard deviation is \$16.80.

(a) What is the data type? **(b)** Write down the values of: p, $\hat{p}$, μ, $\bar{x}$, σ and s. **(c)** Calculate a 99% confidence interval for the average spend. **(d)** What value of n is needed to estimate the true average spend within $\pm$ \$0.50 with 99% confidence? (i.e., $z_{0.005} \frac{s}{\sqrt{n}} = 0.5$)

Question 3

A manufacturer of aircraft parts wishes to estimate the operating life of a particular component. Thus, a sample of 45 components are used until failure. It is found that the average life is 671.23 hours and the variance is 400 hours-squared.

(a) What type of data was collected?

(b) What is the parameter and its value?

(c) What is the statistic and its value?

(d) Calculate / interpret the 99.9% confidence interval.

Question 4

The government are investigating the difference in proportions of people in rural and urban areas in support of a new policy. Researchers collected data on 154 individuals; of these, 38 lived in rural areas and 116 lived in urban areas. It was found that 52.63% of those in rural areas and 60.34% of those in urban areas support the policy.

(a) What is the true difference in proportions? **(b)** Calculate a 90% confidence for this difference and comment on this interval. **(c)** In the sample of 154 individuals, how many of them support the policy? (i.e., 52.63% of 38 plus 60.34% of 116) **(d)** Based on the answer to part (c), estimate the overall proportion of individuals in support of the policy and construct a 90% confidence interval for the true proportion.

Question 5

Consider the following two types of heat sink used for the purposes of CPU cooling:

	Type 1	Type 2
number tested	50	50
mean CPU temperature	40.1	34.8
standard deviation	2.5	1.1

(a) Identify the parameter of interest. **(b)** Calculate a 95% confidence interval for the parameter and comment. **(c)** How large a sample is required to reduce the *margin of error* in the previous confidence interval to ± 0.4? (note: assume that $n_1 = n_2$)

Question 6

A group of 8 computer science students were randomly selected and asked how many hours they spent gaming last week. The average time was found to be 6.4 hours and the standard deviation was 2.2 hours.

(a) Calculate a 95% confidence interval for μ.

(b) Calculate a 99% confidence interval for μ.

Question 7

Guinness set their bottle-filling machine to put 33cl into each bottle. A sample of 5 bottles were selected at random and measured. The volumes in cl were as follows:

34.1	33.5	32.8	33.1	32.5

(a) Calculate the sample mean and standard deviation. **(b)** Calculate a 95% confidence interval. **(c)** Based on the confidence interval, does it appear that the machine is working correctly?

Question 8

A soft drinks company is working on a new recipe for its best-selling drink. The company intends to carry out a study where participants will taste both flavours (current and new) and then answer the question:

"Do you prefer the new flavour?"

It is assumed that the *current* recipe is superior, i.e., that *less than or equal to* 50% of people prefer the new drink ($p \leq 0.5$).

We wish to test the hypothesis that $p \leq 0.5$.

(a) State the null and alternative hypotheses. **(b)** From a sample of 65 people, we find that 43 people prefer the new recipe. Calculate the test statistic and, hence, the p-value.
(c) Based on the evidence, state your conclusion.

Question 8

The average height of a sample of 16 students was 173cm with a variance of 144cm^2. The average height of the Irish population is 169cm.

(i)) Can it be stated at a significance level of 5% that students are on average taller than the population as a whole?

(ii)) What assumption is used to carry out this test? Is this assumption reasonable?

Question 9

A coal-fired power plant is considering two different systems for pollution abatement. The first system has reduced the emission of pollutants to acceptable levels 68% of the time, as determined from 200 air samples. The second, more expensive system has reduced the emission of pollutants to acceptable levels 70% of the time, as determined from 250 air samples. lf the expen sive system is significantly more eilective than the inexpensive system in reducing the pollutants to acceptable levels, then the management of the power plant will install the expensive system.

(i) Which system will be installed if management uses a significance level of 0.05 in making its decision?

(ii) Construct a 95% confidence interval for the difference in the two proportions. Interpret this interval.

Question 10

It is generally assumed that older people are more likely to vote for the Conservatives than younger people. In a survey, 160 of 400 people over 40 and 120 of 400 people under 40 stated they would vote Conservative.

(i) Do the data support this hypothesis at a significance level of 5%?

(ii) Calculate a 95% confidence interval for the difference between the proportion of people over 40 voting Conservative and the proportion of people below 40 voting Conservative.

Question 1

A sample of individuals were randomly assigned one of two diet plans. Over a 6-week period these individuals followed their assigned plan. Their weight loss was recorded at the end of the 6-week period and the results were as follows:

	Plan 1	Plan 2
sample size	42	50
mean weight loss	7.1 lbs	5.2 lbs
variance	$10.1\ \text{lbs}^2$	$16.1\ \text{lbs}^2$

We wish to test the hypothesis that there is no difference between diet plans.

(a) State the null and alternative hypotheses. **(b)** Calculate the test statistic.
(c) Calculate the p-value. **(d)** What is your conclusion?

Question 2

A company claims that it pays men and women equally. The salaries for some randomly selected employees (in thousands) were recorded and the results were as follows:

	Male	Female
sample size	5	3
mean salary	30.2	28.4
standard deviation	1.7	1.9

(a) The F-test was carried out and a p-value of 0.7297 was obtained. What does this mean? **(b)** We wish to test the hypothesis that there is no difference between salaries - what are the null and alternative hypotheses? **(c)** If testing at the 10% level, what is the rejection region? (note your answer to part (a)). **(d)** Is there evidence to suggest gender inequality?

Question 3

A sample of students from two universities was randomly selected. Each student had to complete the same programming task and the time to completion was recorded in each case. The results were as follows:

	University A	University B
sample size	15	15
mean	12.5 hrs	11.1 hrs
variance	3 hrs^2	1.5 hrs^2

We wish to test the hypothesis that there is no difference between universities at the 5% level.

(a) State the null and alternative hypotheses. **(b)** If we do *not* assume equal variances, what are the critical values? **(c)** Calculate the test statistic and, hence, provide your conclusion. **(d)** Between what two values does the p-value lie? (note: the p-value cannot be calculated exactly using the t-tables)

Question 4

The government wish to know if there is a difference in the proportions of people living in rural and urban areas in support of a new policy. From a sample of 38 people in rural areas, it was found that 20 support the policy and from a sample of 116 individuals in urban areas, it was found that 70 support the policy.

(a) Test the hypothesis that there is no difference in proportions at the 5% level.

Question 5

A sample of 100 individuals were asked which product they prefer and the results were as follows:

	Product 1	Product 2	Product 3	Product 4	Product 5	$\sum$
Frequency	19	24	24	14	19	100

(a) If there was no difference between products, what would the expected frequencies be? **(b)** Test the hypothesis that the observed matches the expected (use $\alpha = 0.05$).

Question 6

We wish to test the hypothesis that the following sample has come from a normal distribution:

x	< 5	$5-7$	$7-9$	$9-11$	$11-13$	$13-15$	$15-17$	> 17	$\sum$
Frequency	3	10	23	62	39	14	6	3	160

where the mean and standard deviation were also calculated: $\bar{x} = 10.4$ and $s = 2.5$. With these estimates for μ and σ it is easy to calculate the following probabilities (normal tables):

x	< 5	$5-7$	$7-9$	$9-11$	$11-13$	$13-15$	$15-17$	> 17	$\sum$
p_i	0.015	0.072	0.201	0.307	0.256	0.116	0.029	0.004	1.00

(a) Calculate the expected frequencies (group classes where $e_i < 5$). **(b)** Calculate the test statistic. **(c)** Can we reject the hypothesis that the 5% level? **(c)** Between what two values does the p-value lie?

Question 7

One hundred computer science graduates from each of three different universities were asked how many programming languages they are competent in. The results were as follows: **(a)** Calculate the expected frequencies assuming independence of the two variables.
(b) Calculate the χ^2 statistic and, hence, a range within which the p-value lies.
(c) What is your conclusion? **(d)** Calculate the raw difference scores $(o_i - e_i)$ and comment.