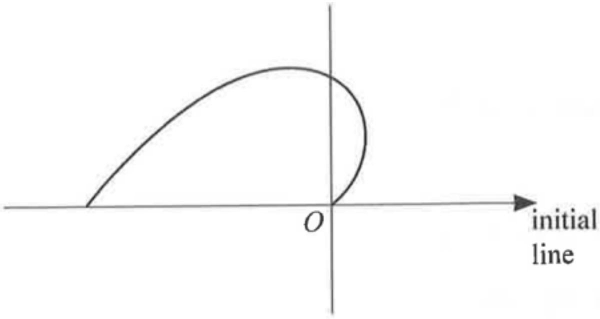
Section A: Pure Mathematics [50 marks]

1 The diagram shows the curve with polar equation $r = 2\ln\theta$ for $1 \ll \theta \ll \pi$.



The area of the region bounded by the curve and the line $\theta = \pi$ is denoted by A. Determine the exact value of A. [6]

1	$A = \frac{1}{2}\int_1^\pi r^2 \, d\theta$ $= \frac{1}{2}\int_1^\pi 4(\ln\theta)^2 \, d\theta$ $= 2\left[\theta(\ln\theta)^2 \Big\|_1^\pi - \int_1^\pi \theta \cdot 2\ln\theta \cdot \frac{1}{\theta} \, d\theta \right]$ $= 2\left[\pi(\ln\pi)^2 - 2\left(\theta\ln\theta \Big\|_1^\pi - \int_1^\pi \theta \cdot \frac{1}{\theta} \, d\theta \right) \right]$ $= 2\left[\pi(\ln\pi)^2 - 2\pi\ln\pi + 2(\pi - 1) \right]$

2 Let $\text{f}(x) = 2^x - x - 3$.

(a) By sketching $y = 2^x$ and $y = x + 3$ on a single diagram, show that the equation $\text{f}(x) = 0$ has exactly two roots. [2]

The two roots of $\text{f}(x) = 0$ are denoted by α and β, where $\alpha < \beta$.

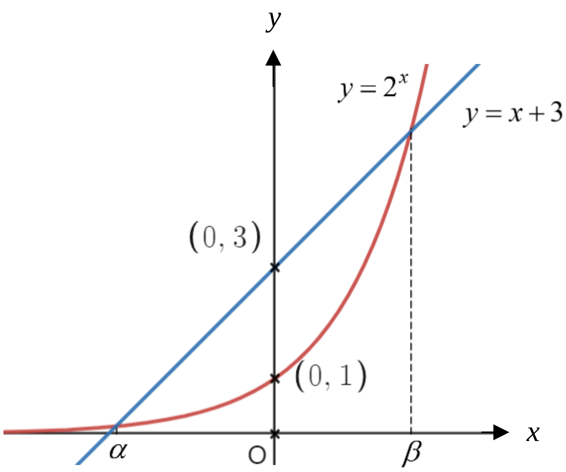
(b) Show that $-3 < \alpha < -2$. [2]

(c) **(i)** Find, in an exact form, the value of x for which $\text{f}(x)$ is a minimum. [2]

(ii) Use the Newton-Raphson method, with initial approximation $x_0 = 0.5$, to calculate the value of x_1 correct to 4 significant figures. [2]

(iii) With reference to your answer to **(c)(i)**, explain why $x_0 = 0.5$ is an unsuitable starting-value for using this iterative method to find an approximation to β. [2]

(iv) Use the Newton-Raphson method to calculate the value of α correct to 4 decimal places. [2]

<table>
<tr><td>(a)</td><td>


$\text{f}(x) = 0 \Rightarrow 2^x = x + 3.$
From the graph, we see that the graphs of $y = 2^x$ and $y = x + 3$ intersect exactly twice. Hence the equation has exactly two real roots.
</td></tr>
<tr><td>(b)</td><td>

x	-3	-2
2^x	$\frac{1}{8}$	$\frac{1}{4}$
$x+3$	0	1

$$\left.\begin{array}{ll} 2^x > x+3, & x=-3 \\ 2^x < x+3, & x=-2 \end{array}\right\} \Rightarrow 2^x = x+3 \text{ for a certain } x \in (-3,-2).$$
</td></tr>
<tr><td>(c)(i)</td><td>
Consider $\text{f}'(x) = 0$.

$$2^x \ln 2 - 1 = 0$$
$$2^x = \frac{1}{\ln 2}$$
$$x = \frac{1}{\ln 2}\ln\left(\frac{1}{\ln 2}\right) = -\frac{1}{\ln 2}\ln(\ln 2)$$

$\text{f}''(x) = 2^x(\ln 2)^2 > 0$ for all x.

$\therefore \text{f}(x)$ is a minimum at $x = -\frac{1}{\ln 2}\ln(\ln 2)$.
</td></tr>
</table>

(ii)	By Newton-Raphson's formula, $x_1 = x_0 - \dfrac{2^{x_0} - x_0 - 3}{2^{x_0}\ln 2 - 1} = -105.2$ (4sf)
(iii)	The minimum point occurs at $x = -\dfrac{1}{\ln 2}\ln(\ln 2) \approx -0.5288$. Hence the tangent to the curve at $x = -0.5$ being near the minimum point has a very gentle slope. As a result, it intersects the x-axis at a faraway point (-105.2,0). $\therefore\ x_0 = 0.5$ is not a suitable starting value.
(iv)	$x_{n+1} = x_n - \dfrac{2^{x_n} - x_n - 3}{2^{x_n}\ln 2 - 1}$ $x_0 = -2$ $x_1 = -2.907207$ $x_2 = -2.862572$ $x_3 = -2.862500$ $x_4 = -2.862500$ $f(-2.86245) = -4.56\times10^{-5}$ $f(-2.86255) = 4.49\times10^{-5}$ $\Rightarrow -2.86255 < \alpha < -2.86245$ $\therefore \alpha = -2.8625$ (4dp)

3 The transformation T is the linear mapping from $\mathbb{R}^4$ to $\mathbb{R}^5$ given by the matrix

$$\mathbf{M} = \begin{pmatrix} 1 & 2 & 1 & 3 \\ 2 & 3 & 1 & 5 \\ 1 & 5 & 1 & -3 \\ 3 & -2 & -1 & 13 \\ -1 & 1 & 2 & 0 \end{pmatrix}.$$

(a) Showing all necessary working, use row operations to find the row rank of $\mathbf{M}$. [5]

(b) **(i)** State the column rank of $\mathbf{M}$, giving a reason for your answer. [1]

(ii) Write down a basis for the range of T. [1]

(c) The kernel of T, ker(T), is defined as the set of vectors $\begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix}$ in $\mathbb{R}^4$ that map to the zero vector in $\mathbb{R}^5$. Determine ker(T). [4]

(a)	$\begin{pmatrix} 1 & 2 & 1 & 3 \\ 2 & 3 & 1 & 5 \\ 1 & 5 & 1 & -3 \\ 3 & -2 & -1 & 13 \\ -1 & 1 & 2 & 0 \end{pmatrix} \xrightarrow[\substack{R_4 \to R_4 - 3R_1 \\ R_5 \to R_5 + R_1}]{\substack{R_2 \to R_2 - 2R_1 \\ R_3 \to R_3 - R_1}} \begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & -1 & -1 & -1 \\ 0 & 3 & 0 & -6 \\ 0 & -8 & -4 & 4 \\ 0 & 3 & 3 & 3 \end{pmatrix}$ $\xrightarrow[\substack{R_4 \to -\frac{1}{4}R_4 \\ R_5 \to \frac{1}{3}R_5}]{\substack{R_2 \to -R_2 \\ R_3 \to \frac{1}{3}R_3}} \begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & -2 \\ 0 & 2 & 1 & -1 \\ 0 & 1 & 1 & 1 \end{pmatrix}$ $\xrightarrow[\substack{R_4 \to R_4 - 2R_2 \\ R_5 \to R_5 - R_2}]{\substack{R_1 \to R_1 - 2R_2 \\ R_3 \to R_3 - R_2}} \begin{pmatrix} 1 & 0 & -1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & -1 & -3 \\ 0 & 0 & -1 & -3 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ $\xrightarrow[R_4 \to R_4 + R_3']{R_3 \to -R_3} \begin{pmatrix} 1 & 0 & -1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ Row rank of **M** = number of non-zero rows in the echelon form = 3
(b)(i)	Column rank = row rank = 3
(ii)	Basis for the range of T= $\left\{ \begin{pmatrix} 1 \\ 2 \\ 1 \\ 3 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \\ 5 \\ -2 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ -1 \\ 2 \end{pmatrix} \right\}$
(c)	$\begin{pmatrix} 1 & 0 & -1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$ $x - z + w = 0$ $y + z + w = 0$ $\left. \begin{array}{l} z + 3w = 0 \Rightarrow z = -3w \\ y = 2w \\ x = -4w \end{array} \right\} \Rightarrow \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = w \begin{pmatrix} -4 \\ 2 \\ -3 \\ 1 \end{pmatrix}$ $\quad \therefore \ker(T) = \left\{ \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} \in \mathbb{R}^4 : \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \alpha \begin{pmatrix} -4 \\ 2 \\ -3 \\ 1 \end{pmatrix} \right\}$

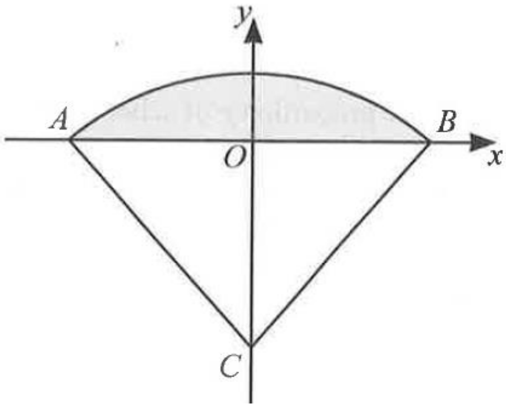
4 A sequence is given by $u_0 = 4$ and $u_{n+1} = (u_n)^2 - 2u_n + 2$ for $n \geq 0$.

By considering $v_n = u_n - 1$, for all $n \geq 0$, find an expression for u_n in terms of n and prove the result by induction. [7]

4	$u_{n+1} = (u_n)^2 - 2u_n + 2$ $= (u_n - 1)^2 + 1$ $= v_n^{\ 2} + 1$ Considering $v_n = u_n - 1$, $\Rightarrow u_n = v_n + 1$ $\therefore u_{n+1} = v_{n+1} + 1$ Comparing with $u_{n+1} = v_n^{\ 2} + 1$, we have $v_{n+1} + 1 = v_n^{\ 2} + 1$ $v_{n+1} = v_n^{\ 2}$ $= \left[(v_{n-1})^2\right]^2$ $= (v_{n-1})^{2^2}$ $= \left[(v_{n-2})^2\right]^4$ $= (v_{n-2})^{2^3}$ $= \ldots$ $= (v_0)^{2^{n+1}}$ $\therefore v_n = (v_0)^{2^n}$ Since $u_n = v_n + 1$, $u_n = (v_0)^{2^n} + 1$. Given $u_0 = 4$ and $v_n = u_n - 1$, $v_0 = u_0 - 1 = 3$ $\therefore u_n = 3^{2^n} + 1$ Let P_n be the proposition that $u_n = 3^{2^n} + 1$ for all $n \geq 0$. For $n = 0$, LHS of $P_0 = u_0 = 4$ (given) RHS of $P_0 = 3^{2^0} + 1 = 3^1 + 1 = 4$ $\therefore P_0$ is true. Assume that P_k is true for some $k \geq 0$, $u_k = 3^{2^k} + 1$ To show P_{k+1} is true also, i.e. $u_{k+1} = 3^{2^{k+1}} + 1$, LHS of P_{k+1}

	$= u_{k+1}$ $= (u_k)^2 - 2u_k + 2$ $= \left(3^{2^k} + 1\right)^2 - 2\left(3^{2^k} + 1\right) + 2$ $= 3^{2^{k+1}} + 2\left(3^{2^k}\right) + 1 - 2\left(3^{2^k}\right) - 2 + 2$ $= 3^{2^{k+1}} + 1$ $=$ RHS of P_{k+1} $\therefore P_k$ is true $\Rightarrow P_{k+1}$ is true Since P_0 is true and P_k is true $\Rightarrow P_{k+1}$ is true, by mathematical induction, P_n is true for all $n \geq 0$

5



The diagram shows a sector *ABC* of a circle with radius *r* and centre *C* on the negative *y*-axis. The angle *ACB* is 2α, where $0<\alpha<\frac{1}{2}\pi$. The shaded region, bounded by the arc *AB* and the *x*-axis, is rotated completely about the *x*-axis to form a solid of revolution, *S*.

(a) (i) By considering the circle with equation $x^2+y^2=r^2$, or otherwise, show that

$$\int_0^{r\sin\alpha}\sqrt{r^2-x^2}\,\mathrm{d}x=\frac{1}{2}\alpha r^2+\frac{1}{2}r^2\sin\alpha\cos\alpha.$$ [2]

(ii) Hence, or otherwise, show that the volume of *S* is $\frac{2}{3}\pi r^3(3\sin\alpha-\sin^3\alpha-3\alpha\cos\alpha)$. [5]

(b) Determine the corresponding expression, in terms of *r* and α, for the surface area of *S*. [7]

(a)(i)	$\int_0^{r\sin\alpha}\sqrt{r^2-x^2}\,\mathrm{d}x=$ area of shaded region $=$ area of sector *OAB* + area of triangle *OBC* $=\frac{1}{2}r^2\alpha+\frac{1}{2}(r\sin\alpha)(r\cos\alpha)$ $=\frac{1}{2}\alpha r^2+\frac{1}{2}r^2\sin\alpha\cos\alpha$ (shown)
(ii)	Equation of arc *AB*: $x^2+(y+r\cos\alpha)^2=r^2$ Volume of *S* $=2\pi\int_0^{r\sin\alpha}y^2\,\mathrm{d}x$ $=2\pi\int_0^{r\sin\alpha}\left(\sqrt{r^2-x^2}-r\cos\alpha\right)^2\mathrm{d}x$ $=2\pi\int_0^{r\sin\alpha}\left(r^2-x^2-2r\cos\alpha\sqrt{r^2-x^2}+r^2\cos^2\alpha\right)\mathrm{d}x$ $=2\pi\left[r^2(1+\cos^2\alpha)x-\frac{1}{3}x^3\right]_0^{r\sin\alpha}-4\pi r\cos\alpha\left(\frac{1}{2}r^2\alpha+\frac{1}{2}r^2\sin\alpha\cos\alpha\right)$ $=2\pi\left[r^3\sin\alpha(1+\cos^2\alpha)-\frac{1}{3}r^3\sin^3\alpha\right]-4\pi r\cos\alpha\left(\frac{1}{2}r^2\alpha+\frac{1}{2}r^2\sin\alpha\cos\alpha\right)$ $=\frac{2}{3}\pi r^3\left[3\sin\alpha(1+\cos^2\alpha)-\sin^3\alpha-3\alpha\cos\alpha-3\sin\alpha\cos^2\alpha\right]$ $=\frac{2}{3}\pi r^3\left[3\sin\alpha-\sin^3\alpha-3\alpha\cos\alpha\right]$ (shown)

(b)	Surface area $= 2\pi \int_0^{r\sin\alpha} y\sqrt{1+\left(\frac{dy}{dx}\right)^2}\,dx$ $x^2 + (y + r\cos\alpha)^2 = r^2$ $\frac{d}{dx}: 2x + 2(y + r\cos\alpha)\frac{dy}{dx} = 0$ $\frac{dy}{dx} = \frac{-x}{y + r\cos\alpha}$ $= 2\pi \int_0^{r\sin\alpha} y\sqrt{1+\left(\frac{dy}{dx}\right)^2}\,dx$ $= 2\pi \int_0^{r\sin\alpha} y\sqrt{1+\left(\frac{-x}{y + r\cos\alpha}\right)^2}\,dx$ $= 2\pi \int_0^{r\sin\alpha} y\sqrt{\frac{r^2}{(y + r\cos\alpha)^2}}\,dx$ $= 2\pi r \int_0^{r\sin\alpha} \frac{y}{y + r\cos\alpha}\,dx$ $= 2\pi r \int_0^{r\sin\alpha} \left(1 - \frac{r\cos\alpha}{y + r\cos\alpha}\right)dx$ $= 2\pi r \int_0^{r\sin\alpha} \left(1 - \frac{r\cos\alpha}{\sqrt{r^2 - x^2}}\right)dx$ $= 2\pi r \left[x - r\cos\alpha \sin^{-1}\frac{x}{r}\right]_0^{r\sin\alpha}$ $= 2\pi r\left[r\sin\alpha - r\alpha\cos\alpha\right]$ $= 2\pi r^2(\sin\alpha - \alpha\cos\alpha)$

Section B: Probability and Statistics [50 marks]

6 An investment analyst models the difference between the actual value of a particular share at a given future time and its predicted value as the continuous random variable X, with the following probability density function.

$$\mathrm{f}(x) = \frac{a}{\pi\left(x^2 + a^2\right)} \quad \text{for all } x \text{ where } a \text{ is a positive constant.}$$

(a) Find in terms of a the value of x_1 for which $\mathrm{P}\left(|X| < x_1\right) = 0.5$. [3]

(b) Show that $\mathrm{E}\left(|X|\right)$ is infinite. [4]

(a)	$\mathrm{P}\left(\lvert X\rvert < x_1\right) = 0.5$ $\mathrm{P}\left(-x_1 < X < x_1\right) = 0.5$ $\Rightarrow \int_{-x_1}^{x_1} \frac{a}{\pi\left(x^2+a^2\right)}\,\mathrm{d}x = 0.5$ $\Rightarrow \frac{2a}{\pi}\int_0^{x_1} \frac{1}{\left(x^2+a^2\right)}\,\mathrm{d}x = 0.5$ $\Rightarrow \frac{2a}{\pi}\left[\frac{1}{a}\tan^{-1}\frac{x}{a}\right]_0^{x_1} = 0.5$ $\Rightarrow \frac{2a}{\pi}\left[\frac{1}{a}\tan^{-1}\frac{x_1}{a} - 0\right] = 0.5$ $\Rightarrow \frac{2}{\pi}\tan^{-1}\frac{x_1}{a} = 0.5$ $\Rightarrow \tan^{-1}\frac{x_1}{a} = \frac{\pi}{4}$ $\Rightarrow \frac{x_1}{a} = 1$ $\Rightarrow x_1 = a$
(b)	$\mathrm{E}\left(\lvert X\rvert\right) = \int_{-\infty}^{\infty} \lvert x\rvert \mathrm{f}(x)\,\mathrm{d}x$ $= \int_0^{\infty} x\mathrm{f}(x)\,\mathrm{d}x + \int_{-\infty}^{0} -x\mathrm{f}(x)\,\mathrm{d}x$ $= 2\int_0^{\infty} x\mathrm{f}(x)\,\mathrm{d}x$ $= \frac{2a}{\pi}\int_0^{\infty} \frac{x}{\left(x^2+a^2\right)}\,\mathrm{d}x$ $= \frac{a}{\pi}\int_0^{\infty} \frac{2x}{\left(x^2+a^2\right)}\,\mathrm{d}x$ $= \frac{a}{\pi}\left[\ln\left(x^2+a^2\right)\right]_0^{\infty}$ As $x \to \infty$, $x^2 + a^2 \to \infty$, $\ln\left(x^2+a^2\right) \to \infty$ $\therefore \frac{a}{\pi}\left[\ln\left(x^2+a^2\right)\right] \to \infty$. Hence $\mathrm{E}\left(\lvert X\rvert\right)$ is infinite.

7 Under usual conditions, on a clear night a meteor is visible on average once every ten minutes. It may be assumed that the event of a meteor being visible is random, and that the length of time for which a meteor is visible is small.

(a) State the two further assumptions needed for the number of meteors that are visible in a fixed period of time to be modelled by a Poisson distribution. [2]

Assume now that a Poisson model is valid, and that usual conditions apply.

(b) Find the probability that, in one hour, between 4 and 10 meteors, inclusive, are visible. [2]

(c) The number of meteors visible in one hour is denoted by M. Find the value of r for which $\mathrm{P}(M=r)=\mathrm{P}(M=r+1)$. [2]

(d) The time in minutes between two consecutive meteors being visible is denoted by T. State the distribution of T, hence find $\mathrm{P}(T>15)$. [4]

(a)	1. Whether one meteor is visible is independent of another meteor being visible. 2. Meteors being visible occurs at constant average rate. Note that in this context, it is not correct to state that meteors occur singly as more than one meteor may be visible at virtually the same time but in different parts of the sky and this does not invalidate the use of a Poisson distribution.
(b)	Let M be the number of meteors visible in one hour. $M \sim \mathrm{P}_0(6)$ $\mathrm{P}(4 \le M \le 10) = \mathrm{P}(M \le 10) - \mathrm{P}(M \le 3)$ $= 0.806175$ ≈ 0.806
(c)	Using GC, $\mathrm{P}(M=5)=\mathrm{P}(M=6)\approx 0.161$ $\therefore r=5$.
(d)	In one min, 0.1 meteor is being visible $\therefore$ T is is an exponential distribution with pdf $\mathrm{f}(t)=\begin{cases} \frac{1}{10}\mathrm{e}^{-\frac{t}{10}}, & t \ge 0 \\ 0 & \text{otherwise} \end{cases}$ $\mathrm{P}(T>15)$ $1-\mathrm{P}(T\le 15)$ $=1-\int_0^{15}\frac{1}{10}\mathrm{e}^{-\frac{t}{10}}\,\mathrm{d}t$ ≈ 0.223 (3sf) (using GC)

8 Jasmine and Kim collect data about the proportion p of school pupils in their district who would support a change in a particular regulation. Jasmine collects a random sample of 50 pupils and finds that the proportion in her sample who would support the change is 28%. Kim independently collects a random sample of 50 pupils and finds the proportion in his sample who would support the change is 36%.

(a) Calculate 90% confidence intervals for p based on

(i) Jasmine's sample,

(ii) Kim's sample.

Give the end-points of the confidence interval correct to 4 significant figures. [3]

(b) Adrian collects an independent random sample of size 100, and uses this sample to calculate an x% confidence interval for p. The limits of this confidence interval are (0.236, 0.484). Calculate the value of x. [4]

(c) By treating Jasmine's sample of 50 pupils and Kim's sample of 50 pupils as a single combined sample of 100 independent observations, obtain a 95% confidence interval for p, based on this combined sample. [2]

(a)	Let $\hat{p}_1$ and $\hat{p}_2$ be the sample proportion of pupils that support a change in a particular regulation collected by Jasmine and Kim respectively.
(i)	Given $\hat{p}_1 = 0.28$ and $\hat{p}_2 = 0.36$, 90% confidence interval for p based on Jasmine's sample is $\left(\hat{p}_1 - 1.6449\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{50}},\quad \hat{p}_1 + 1.6449\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{50}}\right) = (0.1756,\quad 0.3844)$
(ii)	90% confidence interval for p based on Kim's sample is $\left(\hat{p}_2 - 1.6449\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{50}},\quad \hat{p}_2 + 1.6449\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{50}}\right) = (0.2483,\quad 0.4717)$
(b)	Let $\hat{p}_3$ be the sample proportion of pupils that support a change in a particular regulation collected by Adrian. $\left(\hat{p}_3 - z_{\frac{x}{100}}\sqrt{\frac{\hat{p}_3(1-\hat{p}_3)}{100}},\quad \hat{p}_3 + z_{\frac{x}{100}}\sqrt{\frac{\hat{p}_3(1-\hat{p}_3)}{100}}\right) = (0.236,\quad 0.484)$ $\hat{p}_3 = \frac{0.484+0.236}{2} = 0.36 --------(1)$ $\therefore \quad 2z_{\frac{x}{100}}\sqrt{\frac{\hat{p}_3(1-\hat{p}_3)}{100}} = 0.484 - 0.236 --------(2)$ Substitute (1) into (2): $z_{\frac{x}{100}}\sqrt{\frac{0.36(1-0.36)}{100}} = 0.124$ $z_{\frac{x}{100}} = 2.58333$ $\therefore \ x = \text{P}(-2.58333 < Z < 2.58333) \times 100 = 99.021 \approx 99.0$

(c)	Let $\hat{p}_c$ be the sample proportion of the combined sample. $\hat{p}_c = \dfrac{(0.28\times 50)+(0.36\times 50)}{100} = 0.32$ $\therefore$ a 95% confidence interval for *p* based on their single combined sample is $\left(\hat{p}_c - \underbrace{1.960\sqrt{\dfrac{\hat{p}_c(1-\hat{p}_c)}{100}}}_{0.0914},\quad \hat{p}_c + 1.960\sqrt{\dfrac{\hat{p}_c(1-\hat{p}_c)}{100}}\right) = (0.2286,\quad 0.4114)$ (4s.f)

9 An orienteering event was held in 2018 and again in 2019. In 2018 there were a large number of competitors, and the organizer found that the times that competitors took to complete the event were normally distributed with mean 50.0 minutes and standard deviation 17.3 minutes. In 2019 there were fewer competitors. The time, t minutes, to the nearest 0.1 minute, taken by a random sample of 18 competitors to complete the event in 2019 were recorded. The list below gives the value of w, where $w = t - 50$.

–20.3	–17.1	–16.7	–14.4	–12.9	–11.8	3.3	5.6	8.7
13.0	15.5	18.2	19.3	21.8	23.2	24.0	26.4	30.1

First, the organizer carried out a goodness-of-fit test, to investigate whether the times of competitors in 2019 could be considered to have distribution of $N(50.0, 17.3^2)$. The results are shown in the table, in which the decimal entries are given correct to 4 decimal places.

w	$w \le -10$	$-10 < w < 10$	$w \ge 10$
Observed frequency	6	3	9
Probability	0.2816	0.4368	0.2816
Expected frequency	5.0692	7.8617	5.0692
Contribution to test statistic	0.1709	3.0065	3.0481

(a) Explain why it would not have been possible to carry out the goodness-of-fit test using 4 cells rather than 3, no matter what ranges of times had been used. [2]

(b) Show that the goodness-of-fit test suggests, at the 5% significance level, that the distribution $N(50.0, 17.3^2)$ does not give a good fit for the data. [3]

The organiser wishes to test whether the average time for competitors in 2019 was greater than it was in 2018.

(c) Give two reasons why a test based on a normal distribution might be valid, despite the outcome of part **(b)**. [2]

(d) The organiser nevertheless carries out a suitable sign test, at 5% significance level. Determine the conclusion of the test. [5]

(a)	If 4 cells were used, $18 \div 4 = 4.5$, this means that there will be at least one cell with an expected frequency of < 5. Hence it is not possible to carry out the test with 4 cells rather than 3.
(b)	Let T be the time taken by a competitor to complete the event in 2019. H_0 : T follows $N(50.0,\ 17.3^2)$ H_1 : T does not follow $N(50.0,\ 17.3^2)$ Level of significance: 5% Under H_0, From the table, total contribution to test statistic is $0.1709 + 3.0065 + 3.0481 = 6.2255$ Constraint = $\sum O_{ij} = \sum E_{ij}$ Using the chi-square goodness-of-fit test, degree of freedom is $3 - 1 = 2$ Critical region is $\chi^2 > 5.991$

	OR p-value = $\mathrm{P}\left(\chi^2 \geq 6.2255\right) = 0.0445$ Since $6.2255 > 5.991$ or p-value < 0.05 we reject H_0 and conclude that there is sufficient evidence at 5% level of significance that the distribution $\mathrm{N}\left(50.0,\ 17.3^2\right)$ does not give a good fit for the data . Note: The following hypotheses are incorrect: • The data are consistent with N(50.0, 17.32). • The distribution is a good fit for the data. • There is evidence that the distribution is N(50.0, 17.32).
(c)	1. In 2018, the data collected was normally distributed and since the event was only one year later, even with fewer competitors, the time taken by the competitors to complete the event in 2019 should still be normally distributed most likely the same competitors in 2018 might take part in the competition again in 2019. 2. The sample collected in 2019 might be too small which affected the outcome in (b).
(d)	The median time taken by competitors to complete the event in 2018 is 50. Let m be the median time taken by competitors to complete the event in 2019 $\mathrm{H}_0: m = 50$ $\mathrm{H}_1: m > 50$ Level of significance is 5% Let S- be the number of negative signs. i.e number of competitors in 2019 with time taken to complete the event less than the median time taken for competitor in 2018 to complete the event. Under H_0, $S = S_- \sim \mathrm{B}(18, 0.5)$ and observed value of $S = 6$ p-value = $\mathbf{P}(\boldsymbol{S} \leq \mathbf{6})$ = 0.119 > 0.05 Since $p-\text{value} = 0.119 > 0.05$, we do not reject H_0 and conclude that there is insufficient evidence at 5% level that average time for competitors to complete the event in 2019 was greater than it was in 2018.

10 The Human Resources Manager of a large company investigates whether the time, T minutes, taken by employees to travel to work has changed between 2009 and 2019. In 2009 she sampled 80 employees, and in 2019 she sampled 90 employees. The results are recorded in the following table.

Year	n	$\sum t$	$\sum t^2$
2009	80	4976.0	381507.20
2019	90	6516.0	582008.40

The Human Resources Manager carried out a two-sample t-test on the results.

(a) State suitable hypotheses for the test, defining any symbols you use. [2]

(b) Calculate the p-value for the test. [3]

(c) Comment on whether the conclusion of the test can be considered valid, in view of the assumptions needed for the test to be carried out. [3]

(d) Discuss what can be concluded from the test, in the light of your answers to parts **(b)** and **(c)**. [2]

(e) It was subsequently found that the value $\sum t^2 = 381507.20$ had been wrongly recorded. The correct value was greater than 381507.20. Without carrying out any calculations, explain whether the use of the correct value would increase or decrease the strength of the evidence for rejecting the null hypothesis. [2]

<table>
<tr><td>(a)</td><td>Let T_1 and T_2 be the times taken by employees to travel to work in 2009 and 2019 respectively.
Let μ_1 and μ_2 be the <u>population</u> mean times taken by employees to travel to work in 2009 and 2019 respectively.
$\mathrm{H}_0 : \mu_1 - \mu_2 = 0$
$\mathrm{H}_1 : \mu_1 - \mu_2 \neq 0$</td></tr>
<tr><td>(b)</td><td>Two sample t-test was carried out.

Under H_0,
Test statistic:

$T = \dfrac{\left(\overline{T}_1 - \overline{T}_2\right)}{S_p\left(\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}\right)} \sim t_{n_1+n_2-2}$ where $S_p^{\ 2}$ is the pooled variance of the two sample.

Computation,

$\overline{t}_1 = \dfrac{\sum t_1}{n_1} = \dfrac{4976}{80}$ $\overline{t}_2 = \dfrac{\sum t_2}{n_2} = \dfrac{6516}{90}$

$s_1^{\ 2} = \dfrac{1}{80-1}\left(381507.20 - \dfrac{4976^2}{80}\right) = 911.3924051$
$s_2^{\ 2} = \dfrac{1}{90-1}\left(582008.40 - \dfrac{6516^2}{90}\right) = 1238.764045$</td></tr>
</table>

	$s_p{}^2 = \dfrac{(80-1)s_1{}^2 + (90-1)s_2{}^2}{80+90-2} = 32.9366275^2$ Using GC, $p-\text{value} = 2\text{P}\left(T < \dfrac{(\bar{t}_1 - \bar{t}_2)}{s_p\left(\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}\right)}\right) = 0.0455 \ \ (3 \text{ sf})$
(c)	Assumptions needed for two sample *t*-test. 1. Time taken by employees to travel to work in 2009 and 2019 are normally distributed. 2. The two samples collected in 2009 and 2019 are independent of one another. 3. The population variance for the time taken by employees to travel to work in 2009 and 2019 are the same. Assumption 1 might not be valid as T_1 and T_2 need not be normally distributed. But since the sample collected is large enough, by Central Limit theorem, $\bar{T}_1$ and $\bar{T}_2$ are normally distributed approximately. Hence the test statistic $T = \dfrac{(\bar{T}_1 - \bar{T}_2)}{S_p\left(\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}\right)} \sim t_{n_1+n_2-2}$ may still be valid (The t-distribution is close to the standard normal distribution for large values of degrees of freedom.) Assumption 2 is valid as both samples are taken from different year. Assumption 3 may still be valid considering that the company is still located at the same place with similar transport facilities around the area and maybe most of the employees are still working there from 2009 to 2019. Therefore, in all, the conclusion for the test may still be valid to some extent.
(d)	Since $0.01 < p-\text{value} = 0.0455 < 0.05$, H_0 is rejected at 5% level of significance but not rejected at 1% level of significance. Hence, we can conclude that there is sufficient evidence at 5% level of significance that the travel time taken by employees have changed from 2009 to 2019 but not strong enough at 1% level of significance for the same conclusion. In light with some of the assumptions taken to be valid in (c), especially 1 and 3, The *p*-value is close to 0.05 which means that evidence is not very strong to reject H_0 at 5% level of significance.
(e)	Test statistic: $T = \dfrac{(\bar{T}_1 - \bar{T}_2) - (\mu_1 - \mu_2)}{S_p\left(\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}\right)} \sim t_{n_1+n_2-2}$ where $S_p{}^2$ is the pooled variance of the two sample. From the test statistic, only the $S_p{}^2$ is affected by the increase in $\sum t^2$ $s_p{}^2 = \dfrac{(80-1)s_1{}^2 + (90-1)s_2{}^2}{80+90-2}$ Given that $\sum t_1{}^2 > 381507.20$, this implies that $s_1{}^2 = \dfrac{1}{80-1}\left(\sum t_1{}^2 - \dfrac{4976^2}{80}\right) > 911.3924051$

	Therefore $s_p^{\ 2}=\dfrac{(80-1)s_1^{\ 2}+(90-1)s_2^{\ 2}}{80+90-2}>32.9366275^2\ \sum t_1^{\ 2}$ Since $(\bar{t}_1-\bar{t}_2)=-10.2<0$, $s_p^{\ 2}$ increases imply that $\dfrac{(\bar{t}_1-\bar{t}_2)}{s_p\left(\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}\right)}$ will be less negative as compared to when $\sum t^2=381507.20$. This implies that the $p-\text{value}=2\mathrm{P}\left(T<\dfrac{(\bar{t}_1-\bar{t}_2)-(\mu_1-\mu_2)}{s_p\left(\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}\right)}\right)>0.0455$ Higher $p-\text{value}$ indicates decrease in strength of the evidence for rejecting the null hypothesis.