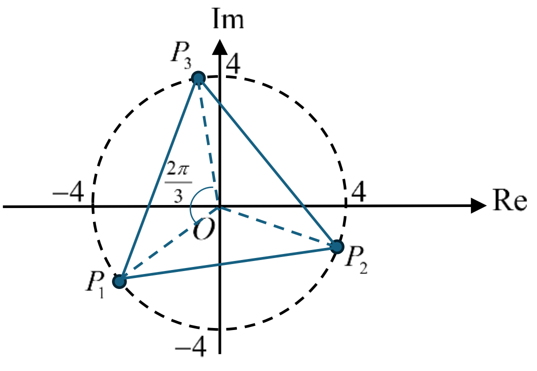
2024 TJC H2 FM Prelim Paper 2 [100 marks]

Section A: Pure Mathematics [50 marks]

1 **(a)** Solve the equation $z^3 = 32 - 32\sqrt{3}\text{i}$, giving your answers z_1, z_2 and z_3 in exact form $r\text{e}^{\text{i}\theta}$, where $r > 0$, $-\pi < \theta \le \pi$ and $\arg z_1 < \arg z_2 < \arg z_3$. [3]

(b) The roots z_1, z_2 and z_3 in part **(a)** are represented by points P_1, P_2 and P_3. Find the exact perimeter of triangle $P_1P_2P_3$. [2]

Solution
(a) $\left\|32-32\sqrt{3}\text{i}\right\| = 32\left\|1-\sqrt{3}\text{i}\right\|$ $= 32\sqrt{1^2+\left(\sqrt{3}\right)^2} = 64$ $\arg\left(32-32\sqrt{3}\text{i}\right) = -\arg\left(1+\sqrt{3}\text{i}\right)$ $= -\frac{\pi}{3}$ $z^3 = 32-32\sqrt{3}\text{i}$ $z^3 = 64\text{e}^{\text{i}\left(-\frac{\pi}{3}+2k\pi\right)}$ $z^3 = 64\text{e}^{\text{i}\frac{\pi}{3}(-1+6k)}$ $\Rightarrow z = 4\text{e}^{\text{i}\frac{\pi}{9}(-1+6k)}$ where $k = 0, \pm 1$ $\Rightarrow z_1 = 4\text{e}^{-\text{i}\frac{7\pi}{9}}$, $z_2 = 4\text{e}^{-\text{i}\frac{\pi}{9}}$, $z_3 = 4\text{e}^{\text{i}\frac{5\pi}{9}}$
(b)  Since the points are equally space apart along the circle, triangle $P_1P_2P_3$ is an equilateral triangle. Length $P_1P_3 = \left\|z_1 - z_3\right\|$ <u>Method 1</u> Using cosine rule

$$|z_1 - z_3|^2 = |z_1|^2 + |z_1|^2 - 2|z_1||z_1|\cos\frac{2\pi}{3}$$

$$= 4^2 + 4^2 - 2(16)\left(-\frac{1}{2}\right)$$

$$= 3(4^2)$$

$$\therefore\ |z_1 - z_3| = 4\sqrt{3}$$

Therefore, perimeter of the triangle is $3\times 4\sqrt{3} = 12\sqrt{3}$

<u>Method 2</u>

$$P_1P_3 = |z_1 - z_3| = \left|4\mathrm{e}^{-\mathrm{i}\frac{\pi}{9}} - 4\mathrm{e}^{\mathrm{i}\frac{5\pi}{9}}\right|$$

$$= \left|4\mathrm{e}^{-\mathrm{i}\frac{\pi}{9}}\right|\left|\left(1 - \mathrm{e}^{\mathrm{i}\frac{6\pi}{9}}\right)\right|$$

$$= 4\left|\mathrm{e}^{\mathrm{i}\frac{2\pi}{6}}\left(\mathrm{e}^{-\mathrm{i}\frac{2\pi}{6}} - \mathrm{e}^{\mathrm{i}\frac{2\pi}{6}}\right)\right|$$

$$= 4\left|\left(-2\mathrm{i}\sin\frac{\pi}{3}\right)\right|$$

$$= 4\sqrt{3}$$

Therefore, perimeter of the triangle is $3\times 4\sqrt{3} = 12\sqrt{3}$

2 The Fibonacci numbers F_n are defined by the conditions $F_0 = 0$, $F_1 = 1$ and

$$F_{n+1} = F_n + F_{n-1} \quad \text{for all } n \geq 1.$$

(a) Compute F_2, F_3, F_4 and F_5. [1]

(b) Compute $F_{n+1}F_{n-1} - F_n^{\ 2}$ for $n = 1, 2, 3, 4$. [2]

(c) Conjecture and prove by induction, for all $n \geq 1$, an expression for $F_{n+1}F_{n-1} - F_n^{\ 2}$. [6]

	Solution
(a)	$F_2 = 1, F_3 = 2, F_4 = 3, F_5 = 5$
(b)	When $n = 1$ $F_2F_0 - F_1^{\ 2} = 1(0) - 1 = -1$ When $n = 2$ $F_3F_1 - F_2^{\ 2} = 2(1) - 1 = 1$ When $n = 3$ $F_4F_2 - F_3^{\ 2} = 3(1) - 2^2 = -1$ When $n = 4$ $F_5F_3 - F_4^{\ 2} = 5(2) - 3^2 = 1$
(c)	Conjecture: $F_{n+1}F_{n-1} - F_n^{\ 2} = (-1)^n$ Let P_n be the statement that $F_{n+1}F_{n-1} - F_n^{\ 2} = (-1)^n$ for $n \in \mathbb{Z}^+$. When $n = 1$, it is proved in part (a) Assume P_k is true for some $k \in \mathbb{Z}^+$, ie. $F_{k+1}F_{k-1} - F_k^{\ 2} = (-1)^k$. When $n = k+1$, $F_{k+2}F_k - F_{k+1}^{\ \ 2} = (F_{k+1} + F_k)(F_{k+1} - F_{k-1}) - F_{k+1}^{\ \ 2}$ $= F_{k+1}^{\ \ 2} + F_kF_{k+1} - F_{k+1}F_{k-1} - F_kF_{k-1} - F_{k+1}^{\ \ 2}$ $= F_kF_{k+1} - F_{k+1}F_{k-1} - F_kF_{k-1}$ $= F_kF_{k+1} \underbrace{-F_{k+1}F_{k-1} + F_k^{\ 2}}_{-(-1)^k} - F_k^{\ 2} - F_kF_{k-1}$ $= F_kF_{k+1} - F_k^{\ 2} - F_kF_{k-1} + (-1)^{k+1}$ $= F_k(F_{k+1} - F_k - F_{k-1}) + (-1)^{k+1}$ $= (-1)^{k+1}$ Since $F_{k+1} = F_k + F_{k-1}$ Alternative solution: When $n = k+1$,

	$F_{k+2}F_k - F_{k+1}{}^2 = (F_{k+1} + F_k)F_k - F_{k+1}{}^2$ $= F_k^2 + F_{k+1}F_k - F_{k+1}(F_k + F_{k-1})$ $= F_k^2 + F_{k+1}F_k - F_{k+1}F_k - F_{k+1}F_{k-1}$ $= F_k^2 - F_{k+1}F_{k-1}$ $= -(F_{k+1}F_{k-1} - F_k^2)$ $= -(-1)^k$ $= (-1)^{k+1}$ Thus P_{k+1} is true. Since P_1 is true, and P_k is true implies P_{k+1} is true. By Mathematical Induction, P_n is true for all positive integer *n*.
	Extension of question: *Explain why 2 consecutive Fibonacci numbers have no common factor greater than 1.* **[Solution]:** $F_{n+1}F_{n-1} - F_n{}^2 = (-1)^n$ Consider the common factor between F_{n-1} and F_n. It divides the LHS of the expression above, this also implies it divides $(-1)^n$. This means the common factor must only be 1. Thus 2 consecutive Fibonacci numbers have no common factor greater than 1.

3 At the start of the month in January 2023, Sandy started her career as a lifestyle vlog content creator with 60 subscribers for her channel. By the end of each month, she lost 5% of her existing subscribers but gained 100 new subscribers each month through her publicity efforts.

Let u_n be the number of subscribers at the end of the n^{th} month taking January 2023 as the first month.

(a) Write down a recurrence relation to model the number of subscribers at the end of the n^{th} month. [2]

(b) Find the general formula of u_n in terms of n. Hence show that the number of subscribers at the end of December 2023 is 952, correct to the nearest integer. [4]

(c) Based on the model in **(a)**, comment on the long-term prospect of Sandy's career. [2]

At the start of January 2024, Sandy decided to collaborate with another well-known content creator to gain more exposure for her own channel. By the end of each month, she gained k% of the number of existing subscribers but lost 100 existing subscribers.

Let w_n be the number of subscribers at the end of the n^{th} month taking January 2024 as the first month.

(d) By considering the general formula of w_n, find an inequality in terms of k that will lead to an increase in the number of subscribers in the long run. Hence find the least integer value of k. [4]

Solution	
(a)	$u_n = 0.95u_{n-1} + 100,\ n \geq 1\ u_0 = 60$
(b)	$u_n = 0.95(0.95u_{n-2} + 100) + 100$ $u_n = 0.95^2 u_{n-2} + 0.95(100) + 100$ $u_n = 0.95^2(0.95u_{n-3} + 100) + 0.95(100) + 100$ $u_n = 0.95^3 u_{n-3} + 0.95^2(100) + 0.95(100) + 100$ $\vdots$ $u_n = 0.95^n u_{n-(n)} + 0.95^{n-1}(100) + 0.95^{n-2}(100) + \cdots + 100$ $u_n = 0.95^n u_0 + 100(1 + 0.95 + 0.95^2 + \cdots + 0.95^{n-1})$ (essential working) $u_n = 0.95^n(60) + 100\left(\dfrac{1-0.95^n}{1-0.95}\right)$ $u_n = 0.95^n(60) + 2000(1-0.95^n)$ $u_n = 2000 - 1940(0.95^n)$

	Alternative method: $u_n = 0.95u_{n-1} + 100$ Let $u_n - \alpha = 0.95(u_{n-1} - \alpha)$ $\alpha - 0.95\alpha = 100 \Rightarrow \alpha = 2000$ $\dfrac{u_n - 2000}{u_{n-1} - 2000} = 0.95$ $u_n - 2000 = (u_0 - 2000)0.95^n$ $u_n = (60 - 2000)0.95^n + 2000$ $u_n = 2000 - 1940(0.95)^n$ At the end of December 2023, $n = 12$ $u_{12} = 2000 - 1940(0.95^{12}) = 951.7014 = 952$ (nearest integer)
(c)	Using the model $u_n = 2000 - 1940(0.95^n)$ as $n \to \infty$, $0.95^n \to 0$ $u_n \to 2000$. From the GC, we observe that the number of subscribers increases slowly and stabilise to 2000 in the long run. Which means she will never have more than 2000 subscribers.
(d)	At the start of January 2024, she should have 952 subscriber from the previous model. $w_n = \left(1 + \dfrac{k}{100}\right)w_{n-1} - 100$, $n \geq 1$ $u_0 = 952$ Using the pattern from (b) $w_n = \left(1 + \dfrac{k}{100}\right)^n u_0 - 100\left(1 + \left(1 + \dfrac{k}{100}\right) + \left(1 + \dfrac{k}{100}\right)^2 + \cdots + \left(1 + \dfrac{k}{100}\right)^{n-1}\right)$ $w_n = \left(1 + \dfrac{k}{100}\right)^n (952) - 100\left(\dfrac{1 - \left(1 + \dfrac{k}{100}\right)^n}{1 - \left(1 + \dfrac{k}{100}\right)}\right)$ $w_n = \left(1 + \dfrac{k}{100}\right)^n (952) + \dfrac{10000}{k}\left(1 - \left(1 + \dfrac{k}{100}\right)^n\right)$ $w_n = \left(1 + \dfrac{k}{100}\right)^n \left(952 - \dfrac{10000}{k}\right) + \dfrac{10000}{k}$ In the long run $\left(1 + \dfrac{k}{100}\right)^n \to \infty$ therefore to have a positive increase

	$952-\frac{10000}{k}>0$ $k>\frac{10000}{952}=10.504$ Alternative method: $w_n=\left(1+\frac{k}{100}\right)w_{n-1}-100$ Consider: $w_n-\beta=r\left(w_{n-1}-\beta\right)$, where $r=1+\frac{k}{100}$ $\beta-r\beta=-100\Rightarrow\beta=\frac{100}{r-1}=\frac{10000}{k}$ $\frac{w_n-\beta}{w_{n-1}-\beta}=r$ $w_n-\beta=\left(w_0-\beta\right)\left(r^n\right)$ $w_n-\frac{10000}{k}=\left(952-\frac{10000}{k}\right)\left(r^n\right)$ $w_n=\left(952-\frac{10000}{k}\right)\left(1+\frac{k}{100}\right)^n+\frac{10000}{k}$ In the long run, $\left(1+\frac{k}{100}\right)^n\to\infty$, therefore to have a positive increase $952-\frac{10000}{k}>0$ $k>\frac{10000}{952}=10.504$ Therefore, least integer value of k is 11.

4 A car travels over a rough surface. The vertical motion of the front suspension is modelled by the differential equation

$$\frac{d^2y}{dt^2} + 25y = 30\cos 5t$$

where y is the vertical displacement of the top of the suspension and t is time.

(a) Find the general solution of the above differential equation. [6]

It is given that $y = 1$ and $\frac{dy}{dt} = 0$ initially.

(b) Find the solution subject to these conditions. [2]

A second model of the motion of the suspension is given by

$$\frac{d^2y}{dt^2} + 2\frac{dy}{dt} + 25y = 30\cos 5t\,.$$

(c) Verify that $y = 3\sin 5t$ is a particular integral for this differential equation. Hence find the general solution. [3]

(d) Compare the behaviour of the suspension predicted by the two models. [1]

Solution	
(a)	A.E. $m^2 + 25 = 0 \quad \Rightarrow m = \pm 5i$ C.F. $y = A\cos 5t + B\sin 5t$ Try P.I. $y = Ct\cos 5t + Dt\sin 5t$ $\frac{dy}{dt} = C\cos 5t + D\sin 5t - 5Ct\sin 5t + 5Dt\cos 5t$ $\frac{d^2y}{dt^2} = -5C\sin 5t + 5D\cos 5t - 5C\sin 5t + 5D\cos 5t$ $-25Ct\cos 5t - 25Dt\sin 5t\,.$ Substituting into $\frac{d^2y}{dt^2} + 25y = 30\cos 5t$, we have $-10C\sin 5t + 10D\cos 5t = 30\cos 5t$ $C = 0, D = 3$ G.S. $y = 3t\sin 5t + A\cos 5t + B\sin 5t$
(b)	When $t = 0, y = 1 \quad A = 1$ $\frac{dy}{dt} = 3\sin 5t + 15t\cos 5t - 5A\sin 5t + 5B\cos 5t$

	Sub $t=0, \frac{\mathrm{d}y}{\mathrm{d}t}=0$, we have $B=0$ P.S. $y=3t\sin 5t+\cos 5t$
(c)	$y=3\sin 5t$ $\frac{\mathrm{d}y}{\mathrm{d}t}=15\cos 5t$ $\frac{\mathrm{d}^2y}{\mathrm{d}t^2}=-75\sin 5t$ Sub into $\frac{\mathrm{d}^2y}{\mathrm{d}t^2}+2\frac{\mathrm{d}y}{\mathrm{d}t}+25y$ We have $-75\sin 5t+2(15\cos 5t)+25(3\sin 5t)=30\cos 5t$ Thus $y=3\sin 5t$ is a particular integral of DE. A.E. $m^2+2m+25=0 \quad m=-1\pm 2\sqrt{6}i$ C.F. $y=\mathrm{e}^{-t}\left(\alpha\cos\sqrt{24}t+\beta\sin\sqrt{24}t\right)$ G.S. $y=\mathrm{e}^{-t}\left(\alpha\cos\sqrt{24}t+\beta\sin\sqrt{24}t\right)+3\sin 5t$
(d)	In the first model, it is not realistic as it predicts that the amplitude of the oscillation will increase without bound in the long run, whereas in the refined model, the amplitude will be at most 3 as time progresses.

5 The ellipse E is given by the equation $\frac{x^2}{9}+\frac{y^2}{8}=1$. $P(x_0, y_0)$ is a point on E where $x_0 > 0$ and $y_0 > 0$.

The tangent to E at P intersects the x-axis at a point Q and O is the origin.

(a) Show that $(\tan\angle POQ)\cdot(\tan\angle PQO)=\frac{8}{9}$. [4]

(b) **(i)** Find the eccentricity of E. [1]

$F_1(-c, 0)$ and $F_2(c, 0)$ are the foci of E, where $c > 0$.

Write down the polar equation of E if the pole is located at

(ii) F_1, [1]

(iii) F_2. [1]

It is given that $\angle PF_2Q=\theta$ and $\angle F_2PQ=\alpha$.

(iv) Using the reflective property of ellipse and both polar equations above, show that $\frac{4}{3+\cos\theta}+\frac{4}{3+\cos(\theta+2\alpha)}=3$. [5]

Solution	**[12 marks]**
(a) $\frac{x^2}{9}+\frac{y^2}{8}=1 \Rightarrow \frac{2x}{9}+\frac{2y}{8}\frac{\mathrm{d}y}{\mathrm{d}x}=0$ Hence $\frac{\mathrm{d}y}{\mathrm{d}x}=-\frac{8x}{9y}$ Gradient of tangent at P = $-\tan\angle PQO$ $\Rightarrow \tan\angle PQO=\frac{8x_0}{9y_0}=\frac{8}{9}\frac{1}{\tan\angle POQ}$ Therefore $(\tan\angle POQ)\cdot(\tan\angle PQO)=\frac{8}{9}$ (shown) (b) (i) $e=\sqrt{1-\frac{b^2}{a^2}}=\sqrt{1-\frac{8}{9}}=\frac{1}{3}$. (ii) The equation is $r=\frac{3\left(1-\frac{1}{9}\right)}{1-\frac{1}{3}\cos\theta}=\frac{8}{3-\cos\theta}$ (iii) The equation is $r=\frac{3\left(1-\frac{1}{9}\right)}{1+\frac{1}{3}\cos\theta}=\frac{8}{3+\cos\theta}$	

(iv) Let QP extended intersect the y-axis at R. By the reflective property of ellipses, $\angle F_1PR = \alpha$.

Then $\angle PF_2Q + (\pi - 2\alpha) = \theta \Rightarrow \angle PF_2Q = (\theta + 2\alpha) - \pi$

Using polar equation in (ii), $F_2P = \dfrac{8}{3+\cos\theta}$

Using polar equation in (iii), $F_1P = \dfrac{8}{3-\cos\left((\theta+2\alpha)-\pi\right)}$

$$= \frac{8}{3+\cos(\theta+2\alpha)}$$

$F_1P + F_2P = 2a \Rightarrow \dfrac{8}{3+\cos\theta} + \dfrac{8}{3+\cos(\theta+2\alpha)} = 2(3) = 6$

Hence $\dfrac{4}{3+\cos\theta} + \dfrac{4}{3+\cos(\theta+2\alpha)} = 3$ (shown)

Section B: Probability and Statistics [50 marks]

6 A random sample of 100 university students were taken from the faculty of Medicine, Dentistry and Health Sciences. The sample consists of 32 Medicine students, 16 Dentistry students and 52 Health Sciences students. Out of the 100 students, 65 were females. A test, at the 1% level of significance was carried out on this data and it was found that there is association between the types of courses enrolled and gender. Given that there were 11 female dentistry students and n male medicine students, find the set of possible values of n. [5]

Solution **[5 marks]**

H_0 : There is no association between types of courses enrolled and gender

H_1 : There is association between types of courses enrolled and gender

	medicine	Dentistry	Health sciences	Total
Female	32 - n	11	22 + n	65
Male	n	5	30 - n	35
Total	32	16	52	100

Under H_0, expected frequency is

	medicine	Dentistry	Health sciences	Total
Female	20.8	10.4	33.8	65
Male	11.2	5.6	18.2	35
Total	32	16	62	100

Degree of freedom = (2 – 1)(3 – 1) = 2

Test Statistic: $\sum \frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} \sim \chi_2^2$

χ^2_{cal}

$$= \frac{(20.8-32+n)^2}{20.8}+\frac{(10.4-11)^2}{10.4}+\frac{(33.8-22-n)^2}{33.8}+\frac{(11.2-n)^2}{11.2}+\frac{(5.6-5)^2}{5.6}+\frac{(18.2-30+n)^2}{18.2}$$

$$= \frac{(n-11.2)^2}{20.8}+\frac{(11.8-n)^2}{33.8}+\frac{(11.2-n)^2}{11.2}+\frac{(n-11.8)^2}{18.2}+\frac{9}{91}$$

For H_0 to be rejected at 1% level of significance,

Calculated value ≥ 9.210

Using GC, when $0 \leq n \leq 5$ or $18 \leq n \leq 30$, calculated value > 9.

Set of possible values of n = $\{n : n \in \mathbb{Z}, 0 \leq n \leq 5 \text{ or } 18 \leq n \leq 30\}$

Extension of question:

A second random sample of size 100N, where N is an integer, is taken from the same faculty. It is found that the proportion of males and females enrolled in the different courses are the same as the first sample. Given that the value of n in the first sample was 16, find the least possible value of N that would lead to the same conclusion (that there is *association between the types of courses enrolled and gender) from a test, at the 1% level of significance, on this second set of data.* *[3]*

Solution:

When $n = 16$,

New calculated value = $N\times\frac{(16-11.2)^2}{20.8}+\frac{(11.8-16)^2}{33.8}+\frac{(11.2-16)^2}{11.2}+\frac{(16-11.8)^2}{18.2}+\frac{9}{91}$

$= 4.75486N$

For H_0 to be rejected at 1% level of significance,

$4.75486N \geq 9.210$

$N \geq 1.94$

Least $N = 2$.

7 The Weibull distribution is a versatile distribution characterised by its ability to model a wide range of data types. It has important applications in meteorological studies such as wind speeds and thunderstorms. The probability density function is given by

$$\mathrm{f}(x) = k\lambda^k x^{k-1} \mathrm{e}^{-(\lambda x)^k}, \quad x > 0$$

where k is the shape parameter and λ is a non-negative scale parameter.

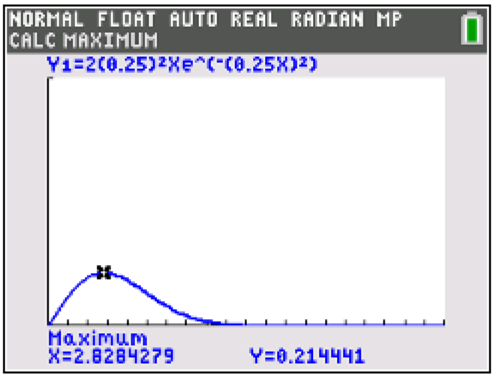
(a) For the case $k = 2$, show that $\mathrm{f}(x)$ is a probability density function for any value of λ. [2]

At a weather station, the wind speed is measured at noon each day. The wind speed, X m/s, is modelled by the Weibull distribution.

(b) For the case $k = 2$ and $\lambda = 0.25$, explain what the shape of the density function indicates about the wind speed at noon at the weather station. Find the most likely wind speed. [2]

The average power output of wind turbines near the weather station is related to the wind speed, x. A model is proposed in which the power, Y megawatts, is given by $Y = CX^3$, where C is a constant.

(c) For the case $k = 2$ and $\lambda = 0.25$, find $\mathrm{E}(Y)$ in terms of C. [2]

	Solution
(a)	When $k = 2$, $\mathrm{f}(x) = 2\lambda^2 x\mathrm{e}^{-(\lambda x)^2}$ $\int_0^\infty \mathrm{f}(x)\,\mathrm{d}x = \int_0^\infty 2\lambda^2 x\mathrm{e}^{-(\lambda x)^2}\,\mathrm{d}x$ $= \left[-\mathrm{e}^{-(\lambda x)^2}\right]_0^\infty$ $= 1$ Recall: if f(x) is a pdf, it must satisfy 1. f(x)≥0 2. $\int_0^\infty \mathrm{f}(x)\,\mathrm{d}x = 1$ Since $\mathrm{f}(x) > 0$ and $\int_0^\infty \mathrm{f}(x)\,\mathrm{d}x = 1$, it is a probability density function.
(b)	When $k = 2$ and $\lambda = 0.25$, $\mathrm{f}(x) = 2(0.25)^2 x\mathrm{e}^{-(0.25x)^2}$.  Lower wind speeds (e.g. less than 5 m/s) are more common than high wind speeds which has a rare chance of occurrence. Using G.C. most likely wind speed is 2.83 m/s.

(c)	Using GC (upper limit set at 20) since the high wind speeds have near zero chance of occurring. $\mathrm{E}(Y)=C\int_0^{\infty} x^3 2(0.25)^2 x\mathrm{e}^{-(0.25x)^2}\ \mathrm{d}x=85.1C$
	Extension of question: *The time between lightning strikes in a thunderstorm, T, can be modelled by the Weibull distribution with* $k=1$ *.* *Show that T has the memoryless property.* [Solution] $\mathrm{f}(x)=\lambda\mathrm{e}^{-(\lambda x)}$ $\mathrm{P}(T>t+c\,\|\,T>c)=\frac{\mathrm{P}(T>t+c)}{\mathrm{P}(T>c)}$ $=\frac{\mathrm{e}^{-\lambda(t+c)}}{\mathrm{e}^{-\lambda c}}$ $=\mathrm{e}^{-\lambda t}$ $=\mathrm{P}(T>t)$ Hence T exhibits memoryless property

8 The Health Promotion Board recommends that adults should engage in at least 150 minutes of moderate-intensity aerobic activity per week. The results of a survey from a random sample of 2105 adults in Singapore show that 1577 adults follow the recommendation.

(a) Calculate a 95% confidence interval for the true proportion p, of adults in Singapore who engage in at least 150 minutes of moderate-intensity aerobic activity per week. Give the end points of the interval correct to 4 significant figures. [2]

(b) Give two reasons why this interval is an approximation. [2]

(c) Suppose that a 90% confidence interval is required, and that the width of the interval is no more than 0.04. Determine the smallest sample size that will satisfy the requirement regardless the value of p. [3]

Solution

(a) $n = 2105,\ \hat{p} = \frac{1577}{2105}$

An approximate 95% confidence interval for p is

$$\left(\hat{p} - 1.9600\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} - 1.9600\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

$= (0.7306499, 0.7676874)$

$= (0.7306, 0.7677)$ (4 s.f.)

$P = \frac{X}{n}$, where X = no. of adults who follow the recommendations.

X is a binomial distribution and when n is large, it is approximated to a normal distribution. Pls note that X is NOT approximated to normal distribution by CLT!

(b) 1) The true proportion p is estimated by $\hat{p}$.

2) The binomial distribution for the number of adults who follow the recommendation out of a random sample of 2105 adults is approximated by a normal distribution.

(c) The width of a 90% confidence interval is $2(1.64485)\sqrt{\frac{p(1-p)}{n}}$.

For $2(1.64485)\sqrt{\frac{p(1-p)}{n}} \le 0.04$,

$$\sqrt{\frac{p(1-p)}{n}} \le 0.012159.$$

Since $p(1-p) = \frac{1}{4} - \left(p - \frac{1}{2}\right)^2 \le \frac{1}{4}$

$$\Rightarrow \sqrt{\frac{p(1-p)}{n}} \le \frac{1}{2\sqrt{n}} \text{ for all values } 0 \le p \le 1$$

$$\Rightarrow \frac{1}{2\sqrt{n}} \le 0.012159$$

$$\Rightarrow \sqrt{n} \ge 41.121$$

$$\Rightarrow n \ge 1690.96$$

$\therefore$ the smallest sample size is 1691.

9 Alvin and Bernard take alternate turns at kicking a football at a goal, and their probabilities of scoring a goal on each kick are p_1 and p_2 respectively, independently of previous outcomes. The first person to score allows the other person one more kick. If the other then scores, the game is drawn. If the other then misses, the first has won the game. Alvin begins a game.

(a) Show that the probability that Alvin scores first, on his *n*th kick, is $p_1\left(q_1q_2\right)^{n-1}$, where $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$. Hence find a simplified expression for the probability that Alvin wins the game. [3]

(b) Find a simplified expression for the probability that the game is drawn. [3]

It is now given that $p_1 = p_2 = \frac{1}{3}$.

Find the expected total number of kicks in the game. [2]

Solution	
(a)	For Alvin to score first in his *n*th kick, both Alvin and Bernard must fail to score in each of their *n*-1 kicks with probability $1-p_1$ and $1-p_2$ respectively and then Alvin scores his *n*th kick with probability p_1. Hence required probability is $p_1\left(q_1q_2\right)^{n-1}$. Probability that Alvin wins $= p_1q_2 + \left(q_1q_2\right)p_1q_2 + \left(q_1q_2\right)^2 p_1q_2 + \ldots$ $= \dfrac{p_1q_2}{1-q_1q_2}$
(b)	Case 1: Game is drawn when Alvin scores first and then Bernard scores with probability = $\dfrac{p_1p_2}{1-q_1q_2}$. Case 2: Game is drawn when Bernard scores first and then Alvin scores with probability $= q_1p_2p_1 + \left(q_1q_2\right)q_1p_2p_1 + \left(q_1q_2\right)^2 q_1p_2p_1 + \ldots$ $= \dfrac{q_1p_2p_1}{1-q_1q_2}$ Hence probability that the game is drawn

$$= \frac{p_1 p_2}{1 - q_1 q_2} + \frac{q_1 p_2 p_1}{1 - q_1 q_2}$$

$$= \frac{p_1 p_2 (1 + q_1)}{1 - q_1 q_2}$$

(c) Let X_A and X_B be the number of kicks made by Alvin and Bernard up to and including the one scored.

$X_A \sim \text{Geo}\left(\frac{1}{3}\right)$ and $X_B \sim \text{Geo}\left(\frac{1}{3}\right)$

$\text{E}\left(X_A\right) = \text{E}\left(X_B\right) = 3$

Expected number of kicks made = 3+1=4

Explanation:

Since $X_A = X_B$, let X be the number of kicks made by either Alvin and Bernard up to and including the one who first score. $X \sim \text{Geo}\left(\frac{1}{3}\right)$

Let Y be the number of kicks to end the game.

Observe that $\text{P}(Y = 2) = \text{P}(X = 1) \times \frac{2}{3}$

> The last person to kick has to lose in order to end the game

$\text{P}(Y = 3) = \text{P}(X = 2) \times \frac{2}{3}$

$\text{P}(Y = 4) = \text{P}(X = 3) \times \frac{2}{3}$

i.e. $Y = X + 1$

Hence $\text{E}(Y) = \text{E}(X + 1) = \text{E}(X) + 1 = 3 + 1$ since $\text{E}(X) = \frac{1}{\frac{1}{3}} = 3$

10 Studies have shown that food rich in a substance known as flavonoid help to lower blood pressure. Flavonoid is found naturally in dark chocolate and is absent in white chocolate. A team of researchers conducted a study to investigate whether there is a greater reduction in blood pressure for people who consume dark chocolate than people who consume white chocolate. A random sample of 19 healthy adults in a particular age group was chosen to participate in the study. Of the 19 healthy adults, 10 were randomly assigned to the dark chocolate group and the rest to the white chocolate group. At the beginning of the study, all participants had their blood pressure recorded in mmHg (millimetres of mercury) before adding a fixed amount of chocolate to their daily diet. At the end of 30 days, their blood pressure was recorded again. The sample mean and standard deviation of the difference of the blood pressures (before diet – after diet) for each group were as follow.

	Sample size	Mean (mmHg)	Standard deviation (mmHg)
Dark	10	6.6	2.9
White	9	4.5	2.6

(a) A researcher concluded that since the difference in the sample means of 2.1 mmHg is greater than 0, there is convincing statistical evidence to conclude that the population mean reduction in blood pressure for those who consume dark chocolate is greater than for those who consume white chocolate. Comment on the researcher's conclusion. [2]

(b) Carry out an appropriate hypothesis test, using a 5% significance level, and state the conclusion that the researcher should reach.

State the assumptions you have made in carrying out the test. [7]

(c) State one non-statistical assumption that is necessary for the validity of the conclusion made in part **(b)**. [1]

(d) One of the values for standard deviation in the table was wrongly computed and the correct value should be greater than shown. Explain whether the use of the correct value would change the conclusion made in part **(b)**. [2]

Solution

(a) The researcher's conclusion may not necessarily be true because looking at the difference in sample means alone does not consider the variability in the sampling distribution of the differences in sample means./

OR

A different random assignment of the 19 participants may lead to different sample means which might in turn lead to a different conclusion.

OR

A difference of 1.6 mmHg could have occurred by random chance, even if the population means are equal.

	AND An appropriate hypothesis test is required to assess the likelihood that the observed difference in sample means occurred by random chance if the population means are equal.
(b)	Let X mmHg and Y mmHg be the difference in the blood pressure reading of a person who consumed dark chocolate and one who consumed white chocolate respectively. Let μ_X mmHg and μ_Y mmHg be the respective population mean difference in blood pressure readings. $H_0: \mu_x = \mu_y$ $H_1: \mu_x > \mu_y$
	Test Statistics: 2-sample t-test When H_0 is true, $T = \dfrac{\left(\bar{X} - \bar{Y}\right)}{S_p\sqrt{\frac{1}{10} + \frac{1}{9}}} \sim t_{17}$ Computation: $\bar{x} = 6.6, \bar{y} = 4.5, \; s_x^2 = \frac{10}{9}(2.9)^2, s_y^2 = \frac{9}{8}(2.6)^2$ $\therefore s_p^2 = \dfrac{9s_x{}^2 + 8s_y{}^2}{17} = \dfrac{10(2.9)^2 + 9(2.6)^2}{17} = 2.9199^2$ $\therefore t = \dfrac{2.1}{2.9199\sqrt{\frac{19}{90}}} = 1.5653$ $p - \text{value} = \text{P}(T > 1.5653) = 0.067970$
	Since p-value > 0.05, H_0 is not rejected at the 5% level of significance. Hence there is insufficient evidence at the 5% significance level to conclude that there is a greater reduction in blood pressure for people who consumed dark chocolate than people who consumed white chocolate. Assumptions: - Differences in blood pressure readings (X and Y) follow normal distributions - Variances of X and Y are equal.
(c)	The rest of the diet / lifestyle are the same for all participants.
(d)	Increase in sample standard deviation will lead to an increase in the pooled sample variance. This will lead to a smaller t value which translates to a bigger p-value. Hence the new p-value $> 0.067970 > 0.05$, H_0 will still be rejected. No change to the conclusion of the test.

11 A dietician is studying the possible effect of a diet plan in lowering cholesterol levels. A random sample of 12 patients is selected and she measures their cholesterol levels before and after they have followed the diet plan for 12 weeks. Their decrease in cholesterol levels measured in millimoles per litre (mmol/L) after following the diet plan for 12 weeks are recorded as follows:

$$-0.1,\ 1.7,\ -1.2,\ 1.1,\ 1.4,\ 0.5,\ 0.9,\ 2.2,\ -1.0,\ 2,\ 0.7,\ 0.3.$$

(a) **(i)** Explain why the dietician has used a paired design. [1]

(ii) Carry out a suitable Wilcoxon matched-pairs signed rank test to examine at the 5% level of significance whether, on the whole, the diet plan is effective in lowering cholesterol level. State any assumption(s) required for the test to be valid. [5]

(iii) The dietician originally planned to carry out a paired sample *t* test. However, she decided that the distributional assumption required for this test might not have been satisfied. State this assumption and explain why in general, if this assumption is satisfied, it is preferable to carry out a *t* test rather than a Wilcoxon test. [2]

(b) Suppose that *k* of the 12 volunteers reported weight loss while the rest of the volunteers reported weight gain after following the diet plan for 12 weeks. Determine the possible values of *k* if a sign test at the 5% level of significance shows that the diet plan had also helped in weight loss. [4]

Solution **[12 marks]**

(a) (i) The pairing will eliminate any differences in individual' s lifestyle/health condition which may affect cholesterol level.

(ii) Let *D* be the decrease in cholesterol level of the 12 volunteers and m_d = population median of *D*

Assumptions: *D* has a continuous and symmetric distribution.

Volunteer	1	2	3	4	5	6	7	8	9	10	11	12
d_i	− 0.1	1.7	−1.2	1.1	1.4	0.5	0.9	2.2	−1.0	2	0.7	0.3
Ranks	1	10	8	7	9	3	5	12	6	11	4	2

$H_0 : m_d = 0$

$H_1 : m_d > 0$

Level of significance: 1%

P = sum of the ranks corresponding to the positive differences

= 10 + 7 + 9 + 3 + 5 + 12 + 11 + 4 + 2 = 63

Q = sum of the ranks corresponding to the negative differences

= 1 + 8 + 6 = 15

$T_{\text{cal}} = Q = 15$

Now, $n = 12$, for a 1-tailed test at 5% level of significance, critical region = $\{t:\ t \le 17\}$

Since T_{cal} falls in the critical region, we reject H_0.

Hence, there is sufficient evidence at the 5% level of significance to conclude that the diet plan is effective in lowering cholesterol level.

(iii) Assume that the population of **differences** in the cholesterol levels before diet and after diet is **normally distributed**.

The *t*-test is more powerful as it takes into account the actual values of the differences, whereas the Wilcoxon tests only utilise the magnitude of the ranks of the differences. Hence the *t*-test is more preferable than the Wilcoxon test.

(b) Let M kg be the population median weight loss. (weight before – weight after)

$H_0: M = 0$

$H_1: M > 0$

Level of significance: 5%

Let S_+ be the number of patients with weight loss

S_- be the number of patients with weight gain

Test Statistic: $S = S_- \sim B\left(12, \frac{1}{2}\right)$ if H_0 is true.

Computation: $S_+ = k$, $S_- = 12 - k$.

Using GC,

k	$P(S \le 12-k)$
9	0.073 > 0.05
10	0.0193 < 0.05
11	0.0032 < 0.05

For H_0 to be rejected, $k \ge 10$

So possible values of k are 10, 11 and 12.