

Statistical Modelling

1 (2 marks)

A statistical model is a simplified mathematical description of a real situation, using probability distributions or relationships between variables **B1**

It is used to make predictions, or to analyse a situation which is too costly, slow or dangerous to investigate directly **B1**

Answer: A simplified mathematical description of a real situation; used to predict or analyse 2 marks

2 (2 marks)

A simple model is easier to use and calculate with, and to explain to others **B1**

A simple model needs fewer parameters to be estimated from data, so it is less likely to be distorted by chance variation in the data **B1**

Answer: Easier to use and interpret; fewer parameters to estimate 2 marks

3 (2 marks)

Emails arrive independently of each other **B1**

Emails arrive at a constant average rate, singly and at random **B1**

Answer: Independent arrivals; constant average rate 2 marks

4 (2 marks)

Each item is defective with the same fixed probability **B1**

The items are independent of each other (the status of one does not affect another) **B1**

Answer: Constant probability; independence 2 marks

5 (3 marks)

$$P(X = 4) = \frac{1}{6} \quad \text{B1}$$

$$E(X) = \frac{7}{2} = 3.5 \quad \text{B1}$$

$$\text{Var}(X) = \frac{35}{12} \text{ (from } E(X^2) = \frac{91}{6} \text{ and } 91/6 - (7/2)^2) \quad \text{B1}$$

Answer: 1/6, 3.5, 35/12 3 marks

6 (3 marks)

(i) Binomial B(15, 0.5) **B1**

(ii) Poisson **B1**

(iii) Exponential **B1**

Answer: (i) binomial, (ii) Poisson, (iii) exponential 3 marks

7 (3 marks)

$$P(X \geq 4) = 1 - P(X \leq 3) \quad \text{M1}$$

$$P(X \leq 3) = 0.4925 \quad \text{A1}$$

$$P(X \geq 4) = 0.5075 \quad \text{A1}$$

Answer: 0.5075 3 marks

8 (4 marks)

$$(i) P(X = 0) = e^{-2.4} = 0.0907 \quad \text{M1 A1}$$

$$(ii) P(X \geq 2) = 1 - P(X = 0) - P(X = 1) \quad \text{M1}$$

$$= 1 - 0.0907 - 0.2177 = 0.6916 \quad \text{A1}$$

Answer: (i) 0.0907, (ii) 0.6916 4 marks

9 (3 marks)

$$z = \frac{185 - 172}{8} = 1.625 \quad \text{M1}$$

$$P(Z > 1.625) = 1 - \Phi(1.625) \quad \text{M1}$$

$$= 0.0521 \quad \text{A1}$$

Answer: 0.0521 3 marks

10 (4 marks)

Mean number of successes = $1 \times 14 + 2 \times 27 + 3 \times 22 + 4 \times 11 + 5 \times 3$ over 80 = 2.413

M1 A1

$5p = 2.413$ so $p = 0.4825$

A1

Expected = $80 \times P(X = 2) = 80 \times 0.3226 = 25.81$

A1

Answer: $p = 0.483$, expected 25.8

4 marks

11 (4 marks)

Total faults = 171, so mean = $\frac{171}{100} = 1.71$

M1 A1

$P(X = 0) = e^{-1.71} = 0.1809$

M1

Expected = $100 \times 0.1809 = 18.1$

A1

Answer: mean 1.71, expected 18.1

4 marks

12 (4 marks)

$P(T > 3) = e^{-0.4 \times 3} = e^{-1.2} = 0.3012$

M1 A1

$P(2 < T < 5) = e^{-0.8} - e^{-2}$

M1

$= 0.4493 - 0.1353 = 0.3140$

A1

Answer: 0.3012 and 0.3140

4 marks

13 (4 marks)

$P(T > 22) = \frac{30 - 22}{30} = \frac{4}{15}$

B1

$E(T) = 15$

B1

$\text{Var}(T) = \frac{30^2}{12} = 75$, so standard deviation = 8.660

M1 A1

Answer: 4/15, 15, 8.66

4 marks

14 (6 marks)

Expected: 20.19, 32.30, 25.84, 13.78, 7.88

M1 A1

$\chi^2 = \sum \frac{(O - E)^2}{E} = 0.509$

M1 A1

$\nu = 5 - 1 - 1 = 3$ (the mean was estimated from the data), critical value 7.815

B1

$0.51 < 7.815$, so there is no evidence to reject the Poisson model

A1

Answer: $\chi^2 = 0.51$, critical 7.815

6 marks

15 (4 marks)

Counters are not replaced, so the probability of red changes after each draw and the trials are not independent

E1

True: $P(2 \text{ red}) = \frac{C(3,2) \times C(7,2)}{C(10,4)} = \frac{3 \times 21}{210} = 0.3000$

M1 A1

Model: 0.2646, which is smaller than the true value

A1

Answer: 0.300 and 0.2646

4 marks

16 (4 marks)

For a Poisson distribution the mean equals the variance

B1

Here the variance is 2.19 times the mean, so the data are over-dispersed and the model is unsuitable

M1 E1

Breakdowns may occur in clusters (not independently), or the rate may vary from week to week

E1

Answer: Variance is more than twice the mean: Poisson unsuitable

4 marks

17 (4 marks)

The class "3 or more" includes all values from 3 upwards, not just 3

E1

Correct probability = $1 - P(X \leq 2) = 1 - 0.7572$

M1

$= 0.2428$

A1

Expected frequency = 24.28

A1

Answer: 24.28

4 marks

18 (6 marks)

$$S_{xx} = 17.50, S_{xy} = 19.60$$

M1 M1

$$b = \frac{S_{xy}}{S_{xx}} = 1.1200$$

A1

$$a = 5.9667 - 1.1200 \times 3.50 = 2.0467$$

A1

$$y = 2.047 + 1.120 \times 4.5 = 7.09$$

M1 A1

Answer: $y = 2.047 + 1.120x$; prediction 7.09 cm

6 marks

19 (5 marks)

$T = A + B$ is normal with mean $12 + 8 = 20$

M1

Variance = $2^2 + 1.5^2 = 6.25$, standard deviation 2.50

M1 A1

$$P(T > 23) = P(Z > \frac{3}{2.50}) = P(Z > 1.20)$$

M1

$$= 0.1151$$

A1

Answer: 0.1151

5 marks

20 (5 marks)

The total is $Po(3.2 + 2.1) = Po(5.3)$

M1 A1

$$P(T \geq 8) = 1 - P(T \leq 7) = 1 - 0.8335$$

M1

$$= 0.1665$$

A1

The sum of two Poisson variables is only Poisson if they are independent

E1

Answer: 0.1665

5 marks

21 (9 marks)

(a) Total goals = 179, mean = $\frac{179}{100} = 1.790$

M1 A1

Answer: 1.790

2 marks

(b) Using $P(X = k)$ with $\lambda = 1.790$: 16.70, 29.89, 26.75, 15.96, 7.14, 3.57

M1

$$\text{Last class: } 100 \times (1 - P(X \leq 4)) = 3.57$$

A1

Answer: 16.70, 29.89, 26.75, 15.96, 7.14, 3.57

2 marks

(c) Classes with expected frequency below 5 are combined, giving expected frequencies 16.70, 29.89, 26.75, 15.96, 10.71

M1

$$\chi^2 = 0.977 \text{ with observed } 14, 31, 30, 15, 10$$

M1 A1

$$\nu = 5 - 1 - 1 = 3, \text{ critical value } 7.815$$

B1

$0.977 < 7.815$: no evidence against the Poisson model

A1

Answer: $0.977 < 7.815$: no evidence against the Poisson model

5 marks

22 (8 marks)

(a) Rate = $\frac{1}{2.5} = 0.40$ per minute

B1

$$P(T > 5) = e^{-0.4 \times 5} = e^{-2} = 0.1353$$

M1 A1

Answer: 0.4, 0.1353

3 marks

(b) $P(T > 5 | T > 3) = \frac{P(T > 5)}{P(T > 3)} = \frac{e^{-2}}{e^{-1.2}}$

M1

$$= e^{-0.8} = 0.4493$$

A1

This equals $P(T > 2)$: the exponential distribution is memoryless, so the time already waited makes no difference

B1 E1

Answer: 0.4493

4 marks

(c) With appointments, arrivals are scheduled rather than random, so a long gap would make an arrival more likely soon, which the memoryless model cannot show

E1

Answer: Scheduled arrivals are not memoryless

1 mark

23 (7 marks)

(a) $X \sim B(200, 0.015)$

B1

$$P(X \leq 2) = 0.4215$$

A1

Answer: 0.4215

2 marks

(b) Mean = $200 \times 0.015 = 3$, so use $Po(3)$

M1

$$P(X \leq 2) = 0.4232$$

Answer: 0.4232

A1

2 marks

$$(c) \text{ Percentage error} = \frac{|0.4215 - 0.4232|}{0.4215} \times 100 = 0.4\%$$

M1 A1

n is large and p is small, so the binomial distribution is well approximated by a Poisson distribution with the same mean

E1

Answer: 0.4%

3 marks

24 (7 marks)

(a) One degree of freedom must also be removed because the mean was estimated from the data

E1

So $\nu = 6 - 1 - 1 = 4$, not 5

E1

Answer: ν should be 4

2 marks

(b) Critical value for $\nu = 4$ at 5% is 9.488

B1

$9.6 > 9.488$ so reject the Poisson model: the data are not consistent with it

M1 A1

Answer: 9.488; reject the Poisson model

3 marks

(c) Failing to reject only shows that the data are consistent with the model, other models could fit equally well, and a small sample gives a test with little power

E1 E1

Answer: Consistency is not proof

2 marks

25 (8 marks)

(a) $\ln y = \ln a + x \ln b$, which is linear in x

B1

Regress $\ln y$ on x : $S_{xx} = 17.50$, $S_{xy} = 6.4739$

M1

$\ln b = 0.3699$ so $b = 1.448$

A1

$\ln a = 3.4004 - 0.3699 \times 2.50 = 2.4756$ so $a = 11.89$

M1 A1

Answer: $a = 11.89$, $b = 1.448$

5 marks

(b) $y = 11.89 \times 1.448^8 = 229$ thousand

M1 A1

8 hours is outside the range of the data (extrapolation) and bacteria cannot keep growing exponentially because of limited resources

E1

Answer: About 229 thousand; extrapolation

3 marks

26 (8 marks)

(a) Total is $Po(4.5)$

M1

$P(T \geq 6) = 1 - P(T \leq 5) = 1 - 0.7029 = 0.2971$

M1 A1

Answer: 0.2971

3 marks

(b) Over four weeks the total is $Po(18)$

M1 A1

$P(T = 20) = 0.0798$

A1

Answer: 0.0798

3 marks

(c) Accidents at the two junctions may be related, for example in bad weather, so are not independent

E1

The accident rate may vary with the season, so is not constant

E1

Answer: Dependence; changing rate

2 marks

27 (11 marks)

$$(a) L(\lambda) = \sum_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = e^{-n\lambda} \lambda^{\sum x_i} / \prod x_i!$$

M1 A1

$$\ln L = -n\lambda + (\sum x_i) \ln \lambda - \ln \prod x_i!$$

A1

$$d(\ln L)/d\lambda = -n + \frac{\sum x_i}{\lambda} = 0, \text{ so } \lambda = \frac{\sum x_i}{n}$$

M1

The second derivative is $-\frac{\sum x_i}{\lambda^2} < 0$, so this is a maximum

A1

Answer: λ -hat = sample mean

5 marks

$$(b) E(\lambda\text{-hat}) = \frac{1}{n} \times n\lambda = \lambda, \text{ so unbiased}$$

M1 A1

$$\text{Var}(\lambda\text{-hat}) = \frac{1}{n^2} \times n\lambda = \frac{\lambda}{n}$$

M1 A1

Answer: Unbiased, variance λ/n

4 marks

(c) Estimate = $\frac{140}{50} = 2.80$

B1

Standard error = $\sqrt{\left(\frac{2.80}{50}\right)} = 0.2366$

B1

Answer: 2.80, 0.237

2 marks

28 (9 marks)

(a) $E(X) = \sum_{x=1}^{\infty} x \frac{e^{-\lambda} \lambda^x}{x!(1 - e^{-\lambda})} = \frac{1}{1 - e^{-\lambda}} \sum_{x=1}^{\infty} x \frac{e^{-\lambda} \lambda^x}{x!}$

M1

The sum is the mean of an ordinary Poisson distribution, which is λ (the $x = 0$ term is zero)

M1 A1

Answer: $E(X) = \frac{\lambda}{1 - e^{-\lambda}}$

3 marks

(b) Equate the mean: $\frac{\lambda}{1 - e^{-\lambda}} = 2.4$, which rearranges to $\lambda = 2.4(1 - e^{-\lambda})$

M1 A1

$\lambda_1 = 2.1823, \lambda_2 = 2.1293, \lambda_3 = 2.1146$

M1

The iteration converges slowly to 2.109

A1

Answer: $\lambda = 2.109$

4 marks

(c) $P(X = 1) = \frac{\lambda e^{-\lambda}}{1 - e^{-\lambda}} = 0.291$

M1 A1

Answer: 0.291

2 marks

29 (11 marks)

(a) $E(X) = 0.6 \times 2 + 0.4 \times 6 = 3.6$

M1 A1

Answer: 3.6

2 marks

(b) For a Poisson variable $E(X^2) = \lambda + \lambda^2$: 6 for Po(2) and 42 for Po(6)

B1

$E(X^2) = 0.6 \times 6 + 0.4 \times 42 = 20.4$

M1 A1

$\text{Var}(X) = 20.4 - 3.6^2 = 7.44$

M1

$7.44 > 3.6$

A1

Answer: $E(X^2) = 20.4, \text{Var}(X) = 7.44$

5 marks

(c) $P(X = 0) = 0.6e^{-2} + 0.4e^{-6} = 0.0822$

M1 A1

Answer: 0.0822

2 marks

(d) A single Poisson distribution has variance equal to mean, but mixing groups with different rates gives a larger variance

E1

The mixture shows how a population of two different groups can look over-dispersed

E1

Answer: Mixing rates gives variance above mean

2 marks

30 (12 marks)

(a) Rate = $\frac{1}{2.1} = 0.4762$

B1

Answer: 0.4762

1 mark

(b) $P(a < T < b) = e^{-ra} - e^{-rb}$ for the finite classes, and $P(T > 5) = e^{-5r}$

M1

Probabilities: 0.3789, 0.2353, 0.1462, 0.1472, 0.0925

M1 A1

Expected frequencies ($\times 100$): 37.89, 23.53, 14.62, 14.72, 9.25

A1

Answer: 37.9, 23.5, 14.6, 14.7, 9.2

4 marks

(c) All expected frequencies are at least 5, so no classes need combining

B1

$\chi^2 = 0.291$

M1 A1

$\nu = 5 - 1 - 1 = 3$, critical value 7.815

B1

$0.29 < 7.815$: no evidence to reject the exponential model

A1

Answer: $0.29 < 7.815$: no evidence to reject the exponential model

5 marks

(d) The model allows arbitrarily long gaps, so the classes must cover all $t > 0$ for the probabilities to total 1

E1

One parameter was estimated from the data, so one further restriction is placed on the expected frequencies

E1

Answer: Classes must cover all values; one parameter estimated

2 marks

Simulation Techniques

1 (2 marks)

A simulation uses a model, usually with random numbers, to imitate a real situation and see what might happen **B1**

Advantage: it can be cheaper, quicker or safer than experimenting on the real situation, and can be repeated many times **B1**

Answer: Imitating a real situation with random numbers; cheap, quick and safe **2 marks**

2 (2 marks)

Changes (such as extra staff) can be tested without disturbing real patients **B1**

Many days of operation can be simulated in a short time **B1**

Answer: Safe testing of changes; speed **2 marks**

3 (2 marks)

Digits 3, 8, 6, 1, 9 give heads for 3 and 1 and tails for the others **M1**

H T T H T **A1**

Answer: H T T H T **2 marks**

4 (2 marks)

Ignore 0, 7 and 9 **M1**

Scores 4, 2, 6, 1 **A1**

Answer: 4, 2, 6, 1 **2 marks**

5 (2 marks)

The event occurs when the number is below 0.35 **M1**

no, no, yes, no, no **A1**

Answer: no, no, yes, no, no **2 marks**

6 (2 marks)

One run of a simulation gives only one random outcome **E1**

Many runs are needed to see the typical behaviour and to reduce the effect of chance variation in the estimates **E1**

Answer: Single run is random; many runs give reliable estimates **2 marks**

7 (3 marks)

00–09 → 0, 10–49 → 1, 50–79 → 2, 80–99 → 3 **B1**

63→2, 08→0, 91→3, 47→1, 22→1 **M1**

Values: 2, 0, 3, 1, 1 **A1**

Answer: 2, 0, 3, 1, 1 **3 marks**

8 (3 marks)

Digits 0–4 → red, 5–7 → blue, 8–9 → green (probabilities 0.5, 0.3, 0.2) **B1**

6→blue, 1→red, 9→green, 3→red, 8→green **M1**

blue, red, green, red, green **A1**

Answer: blue, red, green, red, green **3 marks**

9 (3 marks)

Ignore 9, 0, 8, 7: usable digits 4 2 5 6 1 3 6 6 2 5 **M1**

Pairs: (4, 2), (5, 6), (1, 3), (6, 6) **A1**

Totals: 6, 11, 4, 12 **A1**

Answer: 6, 11, 4, 12 **3 marks**

10 (4 marks)

Cumulative probabilities: 0.2231, 0.5578, 0.8088, 0.9344, 0.9814 **M1 A1**

Ranges: 00–21 → 0; 22–55 → 1; 56–80 → 2; 81–92 → 3; 93–97 → 4; 98–99 → 5 or more **A1**

14, 67, 97 → 0, 2, 4 **A1**

Answer: 0, 2, 4 **4 marks**

11 (5 marks)

Gaps 1, 3, 2, 2 so arrivals at times 1, 4, 6, 8, service times 2, 3, 3, 3

M1 A1

Service starts at 1, 4, 7, 10, so waits are 0, 0, 1, 2

M1 A1

Mean wait = 0.75 minutes

A1

Answer: waits 0, 0, 1, 2, mean 0.75

5 marks

12 (3 marks)

Estimate = $\frac{14}{50} = 0.28$

M1 A1

This is close to 0.3, and the difference is small enough to be explained by chance variation with only 50 runs

E1

Answer: 0.28, close to 0.3

3 marks

13 (3 marks)

$X = 100 + 15Z$

M1

92.20, 119.65, 101.05

A1 A1

Answer: 92.20, 119.65, 101.05

3 marks

14 (4 marks)

Ignore 9 and 0

B1

First six on roll 3, then 2, then 1

M1 A1

Mean = 2.00

A1

Answer: 3, 2, 1, mean 2.00

4 marks

15 (5 marks)

Demands: 2, 3, 1, 0, 2, 1

M1 A1

Unsupplied = demand – 2 when positive: 0, 1, 0, 0, 0, 0

M1 A1

Total = 1

A1

Answer: Demands 2, 3, 1, 0, 2, 1, total unsupplied 1

5 marks

16 (4 marks)

Check: probabilities of the digit groups are 0.5, 0.3, 0.2

B1

Samples: 1,3,2; 3,1,2; 1,2,1

M1 A1

Means: 2.00, 2.00, 1.33

A1

Answer: 2.00, 2.00, 1.33

4 marks

17 (4 marks)

The student's groups contain 26, 35 and 39 numbers, so A has probability 0.26 and C has probability 0.39

M1 E1

A needs 25 numbers, B needs 35 and C needs 40

M1

Correct: A 00–24, B 25–59, C 60–99

A1

Answer: A 00–24, B 25–59, C 60–99

4 marks

18 (5 marks)

Each repetition uses one digit for the coin, then the next allowed digit for the die

B1

H, then 1; T, then 4; T, then 2

M1 A1 A1

Digits not allowed for the die are skipped

A1

Answer: H1; T4; T2

5 marks

19 (4 marks)

Steps: + – + – + – + – + –

M1

Positions: 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0

A1 A1

Final position 0, greatest distance 1

A1

Answer: final 0, greatest distance 1

4 marks

20 (6 marks)

$$\text{Standard error} = \sqrt{\left(\frac{0.32 \times 0.68}{200}\right)} = 0.0330$$

M1 A1

$$\text{Interval} = 0.32 \pm 1.96 \times 0.0330$$

M1

$$(0.255, 0.385)$$

A1

The width is proportional to $\frac{1}{\sqrt{n}}$, so n must be multiplied by 4

M1

800 runs

A1

Answer: (0.255, 0.385); 800 runs

6 marks

21 (7 marks)

(a) Set $U = F(T) = 1 - e^{-0.5T}$

M1

$$e^{-0.5T} = 1 - U, \text{ so } -0.5T = \ln(1 - U)$$

M1

$$T = -2 \ln(1 - U)$$

A1

Answer: $T = -2 \ln(1 - U)$

3 marks

(b) T values: 0.924, 3.430, 0.103, 2.043

M1 A1

Mean = 1.625

A1

Answer: 0.924, 3.430, 0.103, 2.043; mean 1.625

3 marks

(c) For an exponential distribution with rate 0.5 the mean is $\frac{1}{0.5} = 2$, and sample means of simulated values approach the true mean as the sample grows

E1

Answer: True mean is 2

1 mark

22 (7 marks)

(a) Gaps 2, 3, 4, 3, 2, 4 so arrival times 2, 5, 9, 12, 14, 18

M1 A1

Service times 3, 2, 4, 3, 2, 3

B1

Service starts at 2, 5, 9, 13, 16, 18

M1

Waiting times 0, 0, 0, 1, 2, 0

A1

Mean wait = 0.50 minutes

A1

Answer: mean wait 0.50 minutes

6 marks

(b) Only six customers are simulated, so the result has large random variation and the server is not shown running for a long period

E1

Answer: Too few customers

1 mark

23 (7 marks)

(a) Margin = $1.96\sqrt{\left(\frac{p(1-p)}{n}\right)} \leq 0.02$

M1

$$n \geq (98)^2 \times 0.3 \times 0.7 = 2016.8$$

M1

$$n = 2017$$

A1

Answer: 2017 runs

3 marks

(b) The product $p(1-p)$ is at most 0.25 (when $p = 0.5$)

B1

$$n = 98^2 \times 0.25 = 2401$$

A1

Answer: 2401 runs

2 marks

(c) The margin is proportional to $\frac{1}{\sqrt{n}}$

B1

To halve the margin four times as many runs are needed

E1

Answer: Four times the runs to halve the error

2 marks

24 (8 marks)

(a) For $0 < U < 1$, $\ln U$ is negative, so the formula always gives a negative time, which is impossible

E1 E1

Answer: Always negative

2 marks

(b) The rate is $\frac{1}{4} = 0.25$, so $T = -\frac{\ln U}{0.25} = -4 \ln U$

B1

$$U = 0.3 \text{ gives } T = 4.816$$

M1 A1

$$U = 0.8 \text{ gives } T = 0.893$$

A1

Answer: $T = -4 \ln U$: 4.816 and 0.893

4 marks

(c) If U is uniform on $(0, 1)$ then so is $1 - U$

B1

So replacing U by $1 - U$ gives another valid way to produce an exponential value

Answer: $1 - U$ is also uniform

E1

2 marks

25 (7 marks)

(a) $T = -\frac{\ln U}{3}$

B1 B1

Answer: $T = -(\ln U)/3$

2 marks

(b) Gaps: 0.159, 0.350, 0.074, 0.520, 0.199, 0.305

M1 A1

Arrival times (cumulative): 0.159, 0.509, 0.584, 1.104, 1.303, 1.609

M1

3 cars arrive in the first hour

A1

Answer: 3 cars

4 marks

(c) Exponential gaps with rate 3 give a Poisson process, so the number in an hour is $Po(3)$, and a large number of simulated hours would have mean about 3

E1

Answer: $Po(3)$

1 mark

26 (8 marks)

(a) Sums: 1.59, 0.46, 1.59, 1.07, 1.61, 1.15

M1 A1

Sums above 1.5: 3 out of 6, estimate 0.500

A1

Answer: 0.500

3 marks

(b) The region $X + Y > 1.5$ is a triangle with legs 0.5

M1

$$\text{Area} = \frac{1}{2} \times 0.5 \times 0.5 = \frac{1}{8} = 0.125$$

M1 A1

Answer: 0.125

3 marks

(c) The simulated estimate is 0.500 against 0.125, which is a large difference

B1

Six runs is far too few; many more runs are needed for an accurate estimate

E1

Answer: Too few runs

2 marks

27 (11 marks)

(a) $F(x) = \int_0^x 2t \, dt = x^2$

M1 A1

Answer: $F(x) = x^2$

2 marks

(b) $P(\sqrt{U} \leq x) = P(U \leq x^2)$

M1

$$= x^2 \text{ for } 0 \leq x \leq 1, \text{ since } P(U \leq u) = u$$

A1

This equals $F(x)$, so $\sqrt{U}$ has the required distribution

A1

In general $P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$, so $F^{-1}(U)$ has cumulative distribution function F

E1

Answer: $F^{-1}(U)$ has cumulative distribution function F

4 marks

(c) X values: 0.400, 0.700, 0.900, 0.200

M1 A1

Mean = 0.550

A1

$E(X) = \int_0^1 2x^2 \, dx = \frac{2}{3} = 0.667$, the simulated mean 0.550 is a little lower, which is expected with only four values

M1 E1

Answer: 0.40, 0.70, 0.90, 0.20; mean 0.550 against 0.667

5 marks

28 (11 marks)

(a) $f'(x) = 6 - 12x = 0$ gives $x = 0.5$, $f(0.5) = 1.5$

M1 A1

$$P(\text{accept}) = \int_0^1 \frac{f(x)}{1.5} \, dx = \frac{1}{1.5} \times 1 = \frac{2}{3}$$

M1 A1

Answer: Maximum 1.5; $P(\text{accept}) = \frac{2}{3}$

4 marks

(b) The condition is $U_2 \leq 4U_1(1 - U_1)$

B1

Limits: 0.64, 1.00, 0.36, 0.91, 0.96

M1 A1

Compare U_2 with each limit: $0.7 > 0.64$ rejected; $0.95 \leq 1.00$ accepted; $0.5 > 0.36$ rejected; $0.8 \leq$

A1

0.91 accepted; $0.98 > 0.96$ rejected. Accepted values: 0.5, 0.35

Answer: Accepted: 0.5, 0.35

4 marks

(c) Number of pairs = $100 \div \frac{2}{3} = 150$

M1 A1

It only needs values of $f(x)$, not the inverse of F

E1

Answer: 150 pairs**3 marks****29 (12 marks)**(a) $M \leq m$ requires all three dice at most m : $(\frac{m}{6})^3 = \frac{m^3}{216}$ **M1 A1**

$$P(M = m) = \frac{m^3 - (m-1)^3}{216}$$

M1 A1**Answer: $P(M = m) = (m^3 - (m-1)^3)/216$** **4 marks**(b) $P(M = m)$ for $m = 1$ to 6 : 1, 7, 19, 37, 61, 91 over 216**B1** $E(M) = 4.9583$ **M1 A1** $E(M^2) = 25.8935$, so $\text{Var}(M) = 1.3084$ **M1 A1****Answer: $E(M) = 4.958$, $\text{Var}(M) = 1.308$** **5 marks**(c) Standard error = $\sqrt{(\frac{\text{Var}(M)}{n})} \leq 0.05$ **M1**

$$n \geq \frac{1.3084}{0.0025} = 523.4$$

M1 $n = 524$ **A1****Answer: 524****3 marks****30 (11 marks)**(a) Estimate = $\frac{45}{250} = 0.18$ **B1**Standard error = $\sqrt{(\frac{0.18 \times 0.82}{250})} = 0.0243$ **M1 A1** $z = 2.576$ so interval 0.18 ± 0.0626 **M1**

(0.117, 0.243)

A1**Answer: 0.18; (0.117, 0.243)****5 marks**(b) $z\sqrt{(\frac{p(1-p)}{n})} \leq 0.01$ **M1**

$$n \geq (\frac{2.576}{0.01})^2 \times 0.18 \times 0.82$$

M1 $n = 9795$ **A1****Answer: 9795 runs****3 marks**(c) Halving the required error needs four times as many runs, so ± 0.01 costs about 4 times as much as ± 0.02 **M1 E1**

Use a variance reduction method, such as antithetic variables, or a better model of the system

E1**Answer: Four times the cost; variance reduction****3 marks**

Random Number Methods

1 (2 marks) x_0 is the seed (starting value)**B1** m is the modulus**B1****Answer: seed and modulus****2 marks****2 (3 marks)** $x_1 = (5 \times 7 + 3) \bmod 16 = 38 \bmod 16 = 6$ **M1 A1** $x_2 = 1$ and $x_3 = 8$ **A1****Answer: 6, 1, 8****3 marks****3 (2 marks)**

Divide each by 16

M1

0.1875, 0.7500, 0.3125, 0.8750

A1**Answer: 0.1875, 0.7500, 0.3125, 0.8750****2 marks**

4 (2 marks)

The period is the number of values produced before the sequence starts to repeat

B1

With a short period the same values repeat, so they cannot look random in a long simulation

B1

Answer: Length before repeating

2 marks

5 (2 marks)

$1234^2 = 1522756 = 01522756$ as eight digits

M1

Middle four digits: 5227

A1

Answer: 5227

2 marks

6 (2 marks)

They are produced by a fixed rule from the seed, so the sequence is completely determined

E1

They only appear random, and pass tests for randomness

E1

Answer: Deterministic but appear random

2 marks

7 (4 marks)

Values: 7, 0, 1, 4, 2

M1 A1

$x_5 = x_0 = 2$, so the sequence repeats after 5 terms

M1

Period = 5

A1

Answer: period 5

4 marks

8 (4 marks)

Values: 1, 7, 9, 3, 1, 7

M1 A1

It repeats after 4 terms: period 4

A1

Only 4 distinct values can occur out of 10 possible, which is far too few for a useful simulation

E1

Answer: period 4: too short

4 marks

9 (3 marks)

$u = x/16$: 0.750, 0.938, 0.875, 0.562, 0.000

M1

Score = $\text{int}(6u) + 1$: 5, 6, 6, 4, 1

A1 A1

Answer: 5, 6, 6, 4, 1

3 marks

10 (3 marks)

$x = 5 + 7u$

M1

5.84, 9.06, 11.51, 7.17

A1 A1

Answer: 5.84, 9.06, 11.51, 7.17

3 marks

11 (4 marks)

$2916^2 = 8503056 = 08503056$ as eight digits, so the next number is 5030

M1 A1

$5030^2 = 25300900 = 25300900$, so the next number is 3009

M1

$3009^2 = 9054081 = 09054081$, so the next number is 0540

A1

Answer: 5030, 3009, 0540

4 marks

12 (3 marks)

Mean = $\frac{4.95}{10} = 0.495$

M1 A1

The expected mean is 0.5, and the sample mean is close to this

E1

Answer: mean 0.495, expected 0.5

3 marks

13 (3 marks)

Values: 1, 5, 5, 5, 5

M1 A1

The sequence reaches 5 and then stays at 5 for ever, because $4 \times 5 + 1 = 21 = 5 \text{ mod } 16$

E1

Answer: reaches 5 and stays there

3 marks

14 (4 marks)

$c = 3$ and $m = 16$ have no common factor

B1

The only prime factor of 16 is 2, and $a - 1 = 20$ is divisible by 2 and by 4

B1

Sequence: 3, 2, 13, 4, 7, 6, 1, 8, 11, 10, 5, 12, 15, 14, 9, 0

M1

All 16 values from 0 to 15 appear once before the sequence repeats, so the period is 16

A1

Answer: period 16

4 marks

15 (4 marks)

With seed 0 every value is 0

B1

For a different seed the values are all multiples of 4 or fall into a short cycle because the multiplier and modulus share a common factor

E1

Most numbers will never appear, so the numbers are not uniformly spread over the whole range

E1

Testing the output for uniformity and independence is needed, appearance is not enough

E1

Answer: Seed 0 gives all zeros; common factors give short cycles

4 marks

16 (5 marks)

$$T = -\frac{\ln(1-u)}{0.8}$$

M1

Values: 0.248, 1.684, 0.660, 3.324

M1 A1 A1

Mean = 1.479

A1

Answer: 0.25, 1.68, 0.66, 3.32; mean 1.48

5 marks

17 (5 marks)

H_0 : digits are uniform, so each expected frequency is 10

B1

$$\chi^2 = \sum \frac{(O - E)^2}{E} = 6.200$$

M1 A1

$\nu = 9$, critical value 16.919

B1

$6.200 < 16.919$ so there is no evidence of non-uniformity

A1

Answer: $\chi^2 = 6.20$, critical 16.919

5 marks

18 (6 marks)

$n_1 = 10$ above 0.5 and $n_2 = 10$ below; the number of runs is 16

M1 A1

Mean = 11.00, variance = 4.737

M1 A1

$$z = \frac{16 - 11.00}{2.176} = 2.30$$

M1

$|z| > 1.96$ so there is evidence of non-randomness

A1

Answer: 16 runs, $z = 2.30$

6 marks

19 (5 marks)

$$F(x) = \int_0^x t \, dt = \frac{x^2}{2}$$

M1 A1

$$\text{Set } u = \frac{x^2}{2}, \text{ so } x = \sqrt{2u}$$

M1

$u = 0.18 \leq 0.5$, so $x = \sqrt{0.36} = 0.6$

M1 A1

Answer: $x = \sqrt{2u} = 0.6$

5 marks

20 (4 marks)

First four numbers: 0.12, 0.3 are below 0.35, so $X = 2$

M1 A1

Next four: 0.28, 0.19 are below 0.35, so $X = 2$

A1

$E(X) = 4 \times 0.35 = 1.4$

B1

Answer: $X = 2$ and 2; mean 1.4

4 marks

21 (9 marks)

(a) $R = \sqrt{-2 \ln 0.3} = 1.5518$

M1 A1

$\theta = 2\pi \times 0.15 = 0.9425$ radians

B1

$Z_1 = 0.912$, $Z_2 = 1.255$

A1 A1

Answer: $Z_1 = 0.912$, $Z_2 = 1.255$

5 marks

(b) $X = 50 + 4Z$

M1

53.65 and 55.02	A1
Answer: 53.65 and 55.02	2 marks
(c) $Z_1^2 + Z_2^2 = R^2(\cos^2\theta + \sin^2\theta) = R^2$	M1
$= -2 \ln u_1$, which is large only when u_1 is small	E1
Answer: $R^2 = -2 \ln u_1$	2 marks
22 (7 marks)	
(a) $65539 = 2^{16} + 3$	B1
$x_{n+2} \equiv 65539^2 x_n = (2^{32} + 6 \times 2^{16} + 9) x_n$	M1
Since $2^{32} \equiv 0 \pmod{2^{31}}$ this is $(6 \times 2^{16} + 9)x_n = 6(65539 - 3)x_n + 9x_n \equiv 6x_{n+1} - 9x_n$	M1 A1
Answer: $x_{n+2} \equiv 6x_{n+1} - 9x_n \pmod{2^{31}}$	4 marks
(b) Dividing by 2^{31} : $u_{n+2} = 6u_{n+1} - 9u_n + k$ for an integer k	M1
So each triple lies on one of a small number of parallel planes, and is not spread freely through the unit cube, so the numbers are not independent	E1 E1
Answer: Triples lie on a few planes	3 marks
23 (8 marks)	
(a) $c = 1$ and 32 have no common factor	B1
The only prime factor of 32 is 2, $a - 1 = 4$ is divisible by 2 and by 4	B1
Answer: Conditions hold	2 marks
(b) $a = 5$: 1, 6, 31, 28, 13, 2, 11, 24	M1 A1
Period for $a = 5$ is 32	A1
For $a = 3$, $a - 1 = 2$ is not divisible by 4, so the period is less than 32: it is 16	M1 A1
Answer: 32 and 16	5 marks
(c) Full period only means every value appears once per cycle; the sequence of consecutive values can still follow an obvious pattern, such as alternating odd and even values	E1
Answer: Pattern in consecutive values	1 mark
24 (7 marks)	
(a) $x_{16} = 0$ and $x_{17} = 3$	M1 A1
These equal $x_0 = 0$ and $x_1 = 3$: the sequence repeats after 16 values	E1
Answer: $x_{16} = 0$, $x_{17} = 3$: period 16	3 marks
(b) With period 16 only 16 different values are available, repeated in the same order	E1
A simulation needing millions of numbers would reuse the same 16, so the values are neither independent nor suitable	E1
A mean close to the expected value is only one of many properties; passing it does not show the numbers are random	E1
Answer: Period is only 16	3 marks
(c) Use a much larger modulus with multiplier and increment satisfying the full period conditions, for example $m = 2^{32}$	E1
Answer: Larger modulus	1 mark
25 (7 marks)	
(a) $P(X \leq 0) = e^{-2} = 0.1353$	M1 A1
$P(X \leq 1) = 0.4060$, $P(X \leq 2) = 0.6767$, $P(X \leq 3) = 0.8571$, $P(X \leq 4) = 0.9473$, $P(X \leq 5) = 0.9834$	A1
Answer: 0.1353, 0.4060, 0.6767, 0.8571, 0.9473, 0.9834	3 marks
(b) $u = 0.68$: the smallest k with cumulative probability at least 0.68 is 3	M1 A1
$u = 0.13$ gives 0 and $u = 0.95$ gives 5	A1
Answer: 3, 0, 5	3 marks
(c) $X = k$ exactly when $P(X \leq k - 1) < u \leq P(X \leq k)$, an interval of length $P(X = k)$ for a uniform u	E1
Answer: Interval length equals $P(X = k)$	1 mark

26 (7 marks)

(a) Each uniform number has mean $\frac{1}{2}$ and variance $\frac{1}{12}$ B1

The sum has mean $12 \times \frac{1}{2} = 6$ and variance $12 \times \frac{1}{12} = 1$ M1

Subtracting 6 gives mean 0 and variance 1 A1

Answer: Mean 0, variance 1 3 marks

(b) Sum = 6.00 M1

$Z = 6.00 - 6 = 0.00$ A1

Answer: 0.00 2 marks

(c) The central limit theorem B1

Values can never be outside -6 to 6 , so the extreme tails are not modelled correctly E1

Answer: CLT; tails are cut off 2 marks

27 (10 marks)

(a) x_{n+1} depends only on x_n B1

If $x_i = x_j$ with $i < j$ then $x_{i+1} = x_{j+1}$ and so on, so the sequence repeats with period $j - i$ from x_i M1 E1

Answer: Next term depends only on current term 3 marks

(b) Each x_n takes one of the m values $0, 1, \dots, m - 1$ B1

Among the first $m + 1$ terms two must be equal (pigeonhole principle) M1

So a repeat occurs, and by (a) the period is at most m E1

Answer: At most m values 3 marks

(c) If $x_0 = 0$ then $x_1 = 0$, and so on, so every term is 0 M1 A1

$x_{n+1} = ax_n$, so by induction $x_n = a^n x_0 \pmod{m}$ M1 A1

Answer: Period 1; $x_n = a^n x_0$ 4 marks

28 (10 marks)

(a) Set $U = 1 - e^{-(X/\theta)^k}$ M1

$e^{-(X/\theta)^k} = 1 - U$, so $(X/\theta)^k = -\ln(1 - U)$ M1 A1

$X = \theta(-\ln(1 - U))^{1/k}$ A1

Answer: $X = \theta(-\ln(1 - U))^{1/k}$ 4 marks

(b) $X = 3\sqrt{-\ln(1 - U)}$ M1

1.417, 2.498, 4.552 A1 A1

Answer: 1.417, 2.498, 4.552 3 marks

(c) Median = $3\sqrt{\ln 2} = 2.4977$ M1 A1

$F = 1 - e^{-\ln^2} = 1 - \frac{1}{2} = \frac{1}{2}$ M1

Answer: Median 2.498 3 marks

29 (9 marks)

(a) $x_{n+1} = 5x_n + 3 - 16q$ where q is the integer part of $\frac{5x_n + 3}{16}$ M1 A1

So (x_n, x_{n+1}) satisfies $y = 5x + 3 - 16k$ with $k = q$ A1

Answer: $y = 5x + 3 - 16k$ 3 marks

(b) Calculate $k = (5x + 3 - y)/16$ for each of the 16 pairs M1

k takes the values 0, 1, 2, 3, 4 A1

5 different lines are used A1

Answer: $k = 0, 1, 2, 3, 4$: 5 lines 3 marks

(c) If the pairs were independent they would fill the square, but they lie on a few lines, so there is a hidden pattern between consecutive numbers E1 E1

With a large modulus the lines are very close together and there are very many points on them, so the pattern is hard to detect E1

Answer: Pairs lie on few lines 3 marks

30 (11 marks)

(a) For fixed w , $V \leq v$ when U lies in an interval of length v on the circle of numbers modulo 1 M1

Either $U \in [0, v - w] \cup [1 - w, 1)$ when $w \leq v$, giving length $(v - w) + w = v$ M1 A1

or $U \in [1 - w, 1 + v - w)$ when $w > v$, again of length v A1

U is uniform, so the probability equals the length v B1

Answer: Probability is v 5 marks

(b) $P(V \leq v) = \int_0^1 P(V \leq v \mid W = w) g(w) dw = v \int_0^1 g(w) dw$ M1 A1

$= v$, which is the cumulative distribution function of a uniform distribution E1

Answer: V is uniform 3 marks

(c) Sum = 2.22, so the combined value is 0.22 M1 A1

Combined generators have a much longer period, and flaws in one are hidden by the others E1

Answer: 0.22 3 marks

Monte Carlo Simulation

1 (2 marks)

Using random numbers to simulate a large number of trials of a process B1

The results are used to estimate a quantity such as a probability, area or integral B1

Answer: Estimating quantities using many random trials 2 marks

2 (2 marks)

Proportion inside = $\frac{782}{1000}$ estimates $\frac{\pi}{4}$ M1

$\pi \approx 4 \times 0.782 = 3.128$ A1

Answer: 3.128 2 marks

3 (2 marks)

Proportion under the curve = 0.37 M1

Area $\approx 0.37 \times 60 = 22.2$ A1

Answer: 22.2 2 marks

4 (2 marks)

The probability is the area of the quarter circle, $\frac{\pi}{4}$ B1

The proportion of random points inside the circle estimates $\frac{\pi}{4}$, so π is estimated by 4 times the proportion E1

Answer: $\pi/4$ 2 marks

5 (2 marks)

Values: 0.04, 0.36, 0.81 M1

Mean = 0.4033 A1

Answer: 0.403 2 marks

6 (2 marks)

Use more random trials B1

The standard error decreases in proportion to $\frac{1}{\sqrt{n}}$ B1

Answer: Increase the number of trials 2 marks

7 (5 marks)

$x = 2u$: 0.30, 0.90, 1.40, 1.90 M1

x^2 : 0.09, 0.81, 1.96, 3.61, mean = 1.6175 M1 A1

Estimate = interval length $\times$ mean = $2 \times 1.6175 = 3.235$ A1

Exact = $\frac{8}{3} = 2.667$ B1

Answer: 3.235 against exact 2.667 5 marks

8 (4 marks)

$$\text{Estimate} = \frac{31}{200} = 0.155$$

M1 A1

Exact: totals 10, 11, 12 occur in $3 + 2 + 1 = 6$ of 36 ways, $\frac{1}{6} = 0.1667$, so the estimate is a little low but close

M1 A1

Answer: 0.155 against 0.1667

4 marks

9 (3 marks)

$$x^2 + y^2: 0.424, 1.227, 0.269, 1.099, 0.981, 0.476, 0.011, 0.929$$

M1

6 points are inside

A1

$$\pi \approx 4 \times \frac{6}{8} = 3.00$$

A1

Answer: 6 inside, $\pi \approx 3.00$

3 marks

10 (4 marks)

$$p = 0.7854, \text{ so standard error of } p\text{-hat} = \sqrt{\left(\frac{0.7854 \times 0.2146}{1000}\right)} = 0.0130$$

M1 A1

Estimate of π is $4 \times \text{proportion}$, so standard error is $4 \times 0.0130 = 0.0519$

M1 A1

Answer: 0.052

4 marks

11 (4 marks)

$$\text{Estimate} = \frac{14}{40} = 0.35$$

M1 A1

$$\text{Exact} = \int_0^1 x^2 dx = \frac{1}{3} = 0.333$$

M1

The estimate is close to the exact value

A1

Answer: 0.35 against 0.333

4 marks

12 (4 marks)

Mean = 14.06

M1 A1

Sample standard deviation = 1.015

M1

$$\text{Standard error} = \frac{1.015}{\sqrt{10}} = 0.321$$

A1

Answer: mean 14.06, standard error 0.321

4 marks

13 (2 marks)

Hit-or-miss: random points in a rectangle containing the curve, and the proportion below the curve times the rectangle area estimates the integral

B1

Average value: the mean of f at random x values in the interval, times the length of the interval

B1

Answer: Hit-or-miss counts points under the curve; average value uses the mean of f

2 marks

14 (6 marks)

$$x = 1 + 2u: 1.20, 1.70, 2.10, 2.60, 2.90$$

M1

$$\ln x: 0.1823, 0.5306, 0.7419, 0.9555, 1.0647$$

M1 A1

$$\text{Estimate} = 2 \times 0.6950 = 1.390$$

M1 A1

$$\text{Exact} = [x \ln x - x] = 3 \ln 3 - 2 = 1.296$$

B1

Answer: 1.390 against exact 1.296

6 marks

15 (5 marks)

Maxima: 0.72, 0.88, 0.6, 0.45, 0.9, 0.63

M1 A1

Estimate = 0.697

A1

$$P(M \leq m) = m^2, \text{ so } f(m) = 2m \text{ and } E(M) = \int_0^1 2m^2 dm = \frac{2}{3}$$

M1 A1

Answer: estimate 0.697, exact $\frac{2}{3}$

5 marks

16 (6 marks)Proportion = 0.7865 so $\pi \approx 3.1460$

M1 A1

Standard error of proportion = $\sqrt{\frac{0.7865 \times 0.2135}{2000}} = 0.00916$

M1

Interval for π : $4(0.7865 \pm 1.96 \times 0.00916) = (3.074, 3.218)$

M1 A1

The interval contains the true value 3.1416

A1

Answer: 3.146; (3.074, 3.218)

6 marks

17 (4 marks) $e^{0.6} = 1.822$, $e^{1.5} = 4.482$, $e^{2.4} = 11.023$ so the mean is 5.78

M1

The mean of f is only the average height: it must be multiplied by the length of the interval, 3

E1

Correct estimate = $3 \times 5.776 = 17.33$

A1

Exact value = $e^3 - 1 = 19.09$

B1

Answer: 17.33 (times the interval length 3)

4 marks

18 (6 marks) $x^2 + y^2 + z^2$: 0.865, 1.029, 0.580, 1.572, 0.555, 1.074, 0.597, 1.404

M1 A1

4 of the 8 points lie inside the sphere, so the proportion is $\frac{4}{8}$

A1

The cube is one eighth of the sphere, so volume $\approx 8 \times \frac{4}{8} = 4.00$

M1 A1

Exact volume = $\frac{4\pi}{3} = 4.189$

B1

Answer: 4.00 against exact 4.189

6 marks

19 (6 marks) $P(B \text{ or } C \text{ works}) = 1 - 0.2 \times 0.3 = 0.94$

M1 A1

 $P(\text{system}) = 0.9 \times 0.94 = 0.846$

A1

A works if its number is below 0.9; B or C works if one is below 0.8 or 0.7: works, fails, fails, works

M1 A1

Estimate = $\frac{2}{4} = 0.50$

A1

Answer: exact 0.846, simulated 0.50

6 marks

20 (5 marks) $E(U^2) = \frac{1}{3}$ and $E(U^4) = \int_0^1 u^4 du = \frac{1}{5}$

M1 A1

 $\text{Var}(U^2) = \frac{1}{5} - \left(\frac{1}{3}\right)^2 = \frac{4}{45}$

M1 A1

Standard error = $\sqrt{\frac{4}{45 \times 1000}} = 0.0094$

A1

Answer: 4/45; 0.0094

5 marks

21 (7 marks)(a) Each point is under the curve with probability $p = \frac{1}{3}$, so the observation is Bernoulli

M1

Variance = $p(1 - p) = \frac{1}{3} \times \frac{2}{3} = \frac{2}{9} = 0.2222$

A1

Answer: 2/9

2 marks

(b) $\text{Var}(U^2) = E(U^4) - (E(U^2))^2 = \frac{1}{5} - \frac{1}{9}$

M1

 $= \frac{4}{45} = 0.0889$

A1

Answer: 4/45

2 marks

(c) The ratio of variances is $\frac{2}{9} \div \frac{4}{45} = 2.500$

M1 A1

Hit-or-miss needs 2.50 times as many points, so the average value method is more efficient

E1

Answer: 2.50 times

3 marks

22 (8 marks)

(a) $E(e^U) = \int_0^1 e^u du = e - 1$, and $1 - U$ is also uniform, so $E(e^{1-U}) = e - 1$

M1 A1

So $E(Y) = e - 1$, which equals the integral

A1

Answer: $e - 1$

3 marks

(b) $\text{Var}(Y) = \frac{1}{4}[\text{Var}(e^U) + \text{Var}(e^{1-U}) + 2\text{Cov}]$

M1

$= \frac{1}{4}[0.2420 + 0.2420 + 2(-0.2342)] = 0.00391$

M1 A1

Answer: 0.0039

3 marks

(c) $\text{Var}(Y) = 0.0039$ is about 62 times smaller than $\text{Var}(e^U) = 0.2420$

B1

e^U and e^{1-U} are negatively correlated: when one is large the other is small, so the errors partly cancel

E1

Answer: Much smaller variance

2 marks

23 (9 marks)

(a) Proportion crossing $= \frac{255}{400} = 0.637$

M1

$\frac{2}{\pi} = 0.637$ so $\pi = \frac{2}{0.637} = 3.137$

M1 A1

Answer: 3.137

3 marks

(b) Standard error $= \sqrt{\left(\frac{0.637 \times 0.363}{400}\right)} = 0.0240$

M1 A1

Proportion interval (0.5904, 0.6846)

M1

$\pi = \frac{2}{\text{proportion}}: (2.921, 3.388)$

A1

Answer: (2.92, 3.39)

4 marks

(c) It contains the true value 3.1416

B1

π is a reciprocal function of the proportion, which is not linear, so equal changes in the proportion give unequal changes in π

E1

Answer: Reciprocal is non-linear

2 marks

24 (8 marks)

(a) $\hat{p} = \frac{3.18}{4} = 0.795$

M1

Standard error $= 4\sqrt{\left(\frac{0.795 \times 0.205}{500}\right)} = 0.0722$

M1 A1

Answer: 0.072

3 marks

(b) $3.18 - 3.1416 = 0.0384$, which is 0.53 standard errors

M1 A1

This is less than 1.96 standard errors, so 3.1416 lies inside a 95% interval for the estimate

E1

Answer: Within sampling error

3 marks

(c) The student has treated one random estimate as exact: simulations have sampling error

E1

Use many more points, as the standard error falls like $\frac{1}{\sqrt{n}}$

E1

Answer: Sampling error; use more points

2 marks

25 (7 marks)

(a) $X = -\ln u$ (equivalent to $-\ln(1 - u)$)

B1

Values: 0.329, 1.897, 0.916, 0.094

M1 A1

Answer: 0.33, 1.90, 0.92, 0.09

3 marks

(b) $\sqrt{X}$ values: 0.573, 1.377, 0.957, 0.307

M1

Mean = 0.804

A1

Answer: 0.804

2 marks

(c) $\frac{\sqrt{\pi}}{2} = 0.8862$

B1

The estimate 0.804 is close to the exact value; four values give a large standard error

E1

Answer: 0.8862

2 marks

26 (7 marks)**(a)** Proportion inside = 0.3120

M1

Area $\approx 0.312 \times 20 = 6.24$

A1

Answer: 6.24

2 marks

(b) Standard error of the proportion = $\sqrt{\frac{0.312 \times 0.688}{10000}} = 0.00463$

M1 A1

Area: $20(0.312 \pm 1.96 \times 0.00463) = (6.06, 6.42)$

M1 A1

Answer: (6.06, 6.42)

4 marks

(c) Use a smaller rectangle which fits the region more closely, so that a larger proportion of points is inside and the variance of the estimate falls

E1

Answer: Tighter bounding rectangle

1 mark

27 (11 marks)**(a)** $H \sim B(n, p)$ with $p = \frac{A}{R}$

B1

 $E(\hat{A}) = \frac{R}{n} \times np = Rp = A$

M1 A1

Answer: Unbiased

3 marks

(b) $\text{Var}(H) = np(1 - p)$

B1

 $\text{Var}(\hat{A}) = \frac{R^2}{n^2} \times np(1 - p) = \frac{R^2 p(1 - p)}{n}$

M1 A1

 $= \frac{R^2}{n} \times \frac{A}{R} \times \frac{R - A}{R} = \frac{A(R - A)}{n}$

A1

Answer: $\text{Var}(\hat{A}) = \frac{A(R - A)}{n}$

4 marks

(c) $A(R - A)$ is maximised at $A = \frac{R}{2}$

M1 A1

For fixed A, $\text{Var} = \frac{A(R - A)}{n}$ increases with R, so a smaller bounding rectangle gives a smaller variance

M1 E1

Answer: Greatest when $A = R/2$

4 marks

28 (11 marks)**(a)** Hit-or-miss: Bernoulli with probability I, variance $I(1 - I)$

B1

Average value: $\text{Var}(f(U)) = E(f(U)^2) - (E(f(U)))^2$

M1

 $= \int_0^1 f(x)^2 dx - I^2$

A1

Answer: $I(1 - I)$ and $\int_0^1 f(x)^2 dx - I^2$

3 marks

(b) Since $0 \leq f(x) \leq 1$, $f(x)^2 \leq f(x)$

M1

So $\int_0^1 f(x)^2 dx \leq I$

M1

Hence variance of average value = $\int_0^1 f(x)^2 dx - I^2 \leq I - I^2 = I(1 - I)$

A1

Answer: Variance of average value at most $I(1 - I)$

3 marks

(c) Equality needs $f(x)^2 = f(x)$ throughout, so $f(x)$ is 0 or 1 for every x

M1 A1

Answer: f takes only the values 0 and 1

2 marks

(d) Hit-or-miss: $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$, average value: $\frac{1}{3} - \frac{1}{4} = \frac{1}{12}$

M1 A1

Ratio = 3, so 3000 hit-or-miss points are needed

A1

Answer: 1/4 and 1/12; 3000 points

3 marks

29 (12 marks)**(a)** Number exceeding is $B(n, p)$

B1

Standard error = $\sqrt{\frac{p(1 - p)}{n}}$

M1

Relative standard error = $\frac{\sqrt{p(1 - p)}}{n/p} = \sqrt{\frac{1 - p}{np}}$

A1

Answer: $\sqrt{(1 - p)/(np)}$

3 marks

(b) $\frac{1-p}{np} \leq 0.01$ M1

$n \geq \frac{1-0.00135}{0.00135 \times 0.01} = 73974$ M1 A1

Answer: n = 73975 3 marks

(c) Few of the simulated values fall in the rare region, so most of the work is wasted and a huge n is needed for a given relative accuracy E1 E1

Importance sampling places more simulated values in the rare region and corrects with weights E1

Answer: Needs huge n; use importance sampling 3 marks

(d) Weight = $\frac{\text{density of } N(0, 1) \text{ at } y}{\text{density of } N(3, 1) \text{ at } y} = \frac{e^{-y^2/2}}{e^{-(y-3)^2/2}}$ M1

$= e^{-y^2/2 + (y^2 - 6y + 9)/2} = e^{-3y + 4.5}$ M1 A1

Answer: $e^{-3y + 4.5}$ 3 marks

30 (11 marks)

(a) $H \sim B(n, \frac{\pi}{4})$, so $E(4H/n) = 4 \times \frac{\pi}{4} = \pi$ M1 A1

$\text{Var} = \frac{16}{n^2} \times n \times \frac{\pi}{4} (1 - \frac{\pi}{4}) = \frac{\pi(4 - \pi)}{n}$ M1 A1

Answer: Unbiased, variance $\frac{\pi(4 - \pi)}{n}$ 4 marks

(b) $\sqrt{\frac{\pi(4 - \pi)}{10^6}} = 0.00164$ M1 A1

Answer: 0.0016 2 marks

(c) $z = 2.576$ and standard error $\leq \frac{0.001}{2.576}$ M1

$n \geq \frac{\pi(4 - \pi) \times 2.576^2}{0.001^2} = 17895137$ M1 A1

About 17895137 points are needed, a very large number for 3 decimal places, so Monte Carlo is inefficient for low dimensional integrals M1 E1

Answer: 17895137 points 5 marks

Interpreting Statistical Models

1 (2 marks)

12 is the predicted mark with no revision ($x = 0$) B1

3.5 is the predicted increase in mark for each extra hour of revision B1

Answer: 12 marks at $x = 0$; 3.5 marks per hour 2 marks

2 (2 marks)

The difference between an observed value of y and the value predicted by the model B1

Residual = observed – predicted B1

Answer: Observed minus predicted 2 marks

3 (2 marks)

Predicted $y = 2 + 3 \times 12 = 38$ M1

Residual = $41 - 38 = 3$ A1

Answer: 38 and 3 2 marks

4 (2 marks)

The linear model is appropriate for the data B1

The variation is random, with no systematic departure, and the spread is roughly constant B1

Answer: Model appears appropriate 2 marks

5 (2 marks)

The relationship may not stay the same outside the range of the data (extrapolation)

E1

There is no evidence the linear model holds there

E1

Answer: The relationship may change

2 marks

6 (2 marks)

81% of the variation in y is explained by the linear relationship with x

B1

The remaining 19% is unexplained, due to other factors or random variation

B1

Answer: 81% of variation explained

2 marks

7 (3 marks)

Gradient: each 1°C rise in temperature increases sales by about 80 ice creams

B1

Intercept: sales of -420 at 0°C is impossible, so this has no practical meaning

B1 E1

Answer: 80 more sales per °C; intercept not meaningful

3 marks

8 (4 marks)

Predicted values: 5.40, 9.60, 11.70, 15.90, 20.10

M1 A1

Residuals: 0.20, 0.30, -0.70, 0.30, -0.60

A1

At x = 7 the residual is 0.30, so the model underestimates

A1

Answer: residuals 0.20, 0.30, -0.70, 0.30, -0.60

4 marks

9 (4 marks)

Model A residuals: 1.60, 1.10, 0.70, 0.80

M1

$RSS_A = 4.900$

A1

$RSS_B = 5.860$

A1

Model A has the smaller RSS so fits better

A1

Answer: RSS 4.90 and 5.86: Model A

4 marks

10 (3 marks)

The residuals are positive, then negative, then positive: a curved (U-shaped) pattern

M1 A1

This suggests the relationship is not linear, so a curve such as a quadratic should be tried

E1

Answer: Curved pattern: not linear

3 marks

11 (4 marks)

x = 35: y = 20 + 1.5 × 35 = 72.5

M1 A1

x = 80: y = 140.0

A1

x = 35 is within the data range (interpolation) so is more reliable than x = 80 (extrapolation)

E1

Answer: 72.5 and 140.0: x = 35 more reliable

4 marks

12 (2 marks)

The p-value is large, so there is no evidence to reject the Poisson model

B1

The data are consistent with the model, but this does not prove it is correct

E1

Answer: No evidence against the model

2 marks

13 (4 marks)

Standard error = $\frac{2.5}{\sqrt{25}} = 0.5$

M1

Critical value $t_{24} = 2.064$

B1

Interval = $8.4 \pm 2.064 \times 0.5 = (7.37, 9.43)$

M1 A1

Answer: (7.37, 9.43)

4 marks

14 (7 marks)

(a) $S_{xx} = 17.50$, $S_{xy} = 17.70$

M1

b = 1.0114

A1

a = $4.900 - 1.0114 \times 3.50 = 1.3600$

A1

Answer: a = 1.360, b = 1.011

3 marks

(b) Residuals: -0.071, -0.283, 0.506, -0.206, 0.383, -0.329

M1 A1

Sum = 0.000

B1

The residuals have no clear pattern, with a mix of positive and negative values, so the linear model is reasonable

E1

Answer: Sum is zero; no pattern

4 marks

15 (4 marks)

$$y = e^{0.5 + 0.2x} = e^{0.5} \times (e^{0.2})^x$$

M1

$$A = e^{0.5} = 1.649 \text{ and } b = e^{0.2} = 1.221$$

A1 A1

The population grows by a factor of 1.221 each day, an increase of about 22.1%

E1

Answer: A = 1.65, b = 1.221: 22.1% growth per day

4 marks

16 (4 marks)

$$\text{Increase} = 8.9 - 6.2 = 2.7 \text{ minutes}$$

M1

$$\text{Percentage change} = \frac{2.7}{6.2} \times 100 = 43.5\%$$

A1

A 10% rise in calls gives about a 44% rise in waiting time, so the model is very sensitive to the call rate

E1

The call rate must be estimated accurately for the results to be reliable

E1

Answer: 43.5% increase: sensitive to the call rate

4 marks

17 (4 marks)

$r = 0.8$ does not mean 80% of the points are close to the line: it measures the strength of the linear association

E1

The proportion of variation explained is $r^2 = 0.64$, so 64% not 80%

M1 A1

A high correlation does not show that a linear model is appropriate or that x causes y

E1

Answer: $r^2 = 0.64$; r is not a percentage of points

4 marks

18 (7 marks)

$$(a) S_{xx} = 28, S_{xy} = 44.200$$

M1 A1

$$b = 1.5786$$

A1

Answer: 1.579

3 marks

$$(b) \text{ Using the first six points: } b = 0.9914$$

M1 A1

The gradient falls from 1.579 to 0.991, so the point is influential and probably an outlier: it should be checked before being kept

A1 E1

Answer: 0.991; the point is influential

4 marks

19 (5 marks)

$$\text{Linear: } \frac{120}{10} = 12.00$$

M1 A1

$$\text{Quadratic: } \frac{112}{9} = 12.444$$

M1 A1

The quadratic model has the larger variance estimate, so its small gain in RSS does not justify the extra parameter: prefer the linear model

E1

Answer: 12 and 12.44: prefer the linear model

5 marks

20 (4 marks)

$$\text{Expected number} = 40 \times 0.0455 = 1.82$$

M1 A1

The observed 3 is close to the expected value of about 2, so the normality assumption appears reasonable

E1

A normal probability plot would give further support

E1

Answer: About 1.8 expected: reasonable

4 marks

21 (9 marks)

$$(a) S_{xx} = 1050.0, S_{xy} = 957.75$$

M1

$$b = 0.9121$$

A1

$$a = 29.863 - 0.9121 \times 27.5 = 4.779$$

A1

Answer: $y = 4.78 + 0.912x$

3 marks

$$(b) S_{yy} = 885.44, r = \frac{S_{xy}}{\sqrt{(S_{xx}S_{yy})}} = 0.9933$$

M1 A1

$r^2 = 0.9866$: about 98.7% of the variation in sales is explained by the advertising spend	E1
Answer: $r = 0.993$, $r^2 = 0.987$	3 marks
(c) $x = 28$: $y = 4.779 + 0.912 \times 28 = 30.32$	M1 A1
80 is far outside the data range (10 to 45), so the prediction would be an extrapolation and the linear relationship may not continue	E1
Answer: 30.3; 80 is outside the range	3 marks
22 (9 marks)	
(a) Expected frequencies: 20.85, 36.49, 31.93, 18.63, 12.10	M1 A1
$\chi^2 = 0.978$	M1 A1
Answer: $\chi^2 = 0.98$	4 marks
(b) $p = \frac{1}{1 + 1.750} = 0.3636$, expected: 43.64, 27.77, 17.67, 11.25, 19.68	M1 A1
$\chi^2 = 27.351$	A1
Answer: $\chi^2 = 27.35$	3 marks
(c) Critical value 7.815: Poisson gives 0.98 (acceptable) and geometric gives 27.35 (rejected)	A1
The Poisson model has the smaller χ^2 , so it fits better	E1
Answer: Poisson better	2 marks
23 (7 marks)	
(a) The spread of the residuals increases as the fitted value increases, so the variance is not constant	M1 A1
Answer: Non-constant variance	2 marks
(b) Standard errors and confidence intervals assume constant variance, so intervals for large values of y would be too narrow and those for small values too wide	E1 E1
Answer: Intervals would be wrong	2 marks
(c) Take logarithms of y	B1
If the errors are proportional to the size of y , taking logs makes the spread approximately constant	E1
The ratio 1.2 : 2.5 : 4.9 is roughly proportional to the typical size of y	B1
Answer: Log transform of y	3 marks
24 (8 marks)	
(a) Mean = 12.54	B1
$s = 1.037$	M1 A1
Answer: mean 12.54, $s = 1.037$	3 marks
(b) Standard error = $\frac{1.037}{\sqrt{10}} = 0.3280$	M1
$t_9 = 2.262$	B1
Interval = (11.80, 13.28)	A1
Answer: (11.80, 13.28)	3 marks
(c) The mean of the runs is only an estimate of the true mean, and varies from one set of runs to another	E1
The interval shows how uncertain the estimate is: more runs would narrow it	E1
Answer: Estimate, not exact	2 marks
25 (9 marks)	
(a) $C'(q) = -\frac{DS}{q^2} + \frac{h}{2} = 0$, so $q^2 = \frac{2DS}{h} = 60000$	M1 A1
$q = 244.9$, $C = 489.9$	A1
Answer: $q = 244.9$, cost 489.9	3 marks
(b) New $q = \sqrt{\left(\frac{2 \times 1440 \times 50}{2}\right)} = 268.3$	M1 A1
Cost with old q : 538.9, against the best cost of 536.7	A1
Answer: $q = 268.3$; old batch costs 538.9	3 marks
(c) Batch size changes by 9.5%; excess cost = $\frac{538.9 - 536.7}{536.7} \times 100 = 0.42\%$	M1 A1
The cost is not very sensitive to errors in demand, so the model is robust	E1

Answer: 9.5% and 0.42%**3 marks****26 (8 marks)****(a)** Errors: 2, -3, 3, -3, 3, -1**M1 A1**

Mean error = 0.167

A1**Answer: mean error 0.17****3 marks****(b)** $MAE = \frac{15}{6} = 2.500$ **M1 A1** $RMSE = \sqrt{\left(\frac{41}{6}\right)} = 2.614$ **M1 A1****Answer: MAE 2.50, RMSE 2.61****4 marks****(c)** A model can be tuned to fit the data it was built from, including its random variation, and so look better than it will in use**E1****Answer: Avoids over-fitting****1 mark****27 (10 marks)****(a)** $\partial S / \partial a = -2 \sum_{i=1}^n e_i = 0$ **M1 A1** $\partial S / \partial b = -2 \sum_{i=1}^n x_i e_i = 0$ **M1 A1****Answer: $\sum_{i=1}^n e_i = 0$ and $\sum_{i=1}^n x_i e_i = 0$** **4 marks****(b)** $\sum_{i=1}^n e_i = 0$ gives $\sum_{i=1}^n y_i - na - b \sum_{i=1}^n x_i = 0$ **M1**Dividing by n: $\bar{y} = a + b\bar{x}$ **A1****Answer: $\bar{y} = a + b\bar{x}$** **2 marks****(c)** $RSS = \sum_{i=1}^n e_i(y_i - \bar{y} - b(x_i - \bar{x}))$ using the line through $(\bar{x}, \bar{y})$ **M1**The terms in $e_i(x_i - \bar{x})$ sum to zero by (a), so $RSS = S_{yy} - bS_{xy} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$ **M1 A1**

$$\frac{RSS}{S_{yy}} = 1 - \frac{S_{xy}^2}{S_{xx}S_{yy}} = 1 - r^2$$

A1**Answer: $RSS = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$, $r^2 = 1 - \frac{RSS}{S_{yy}}$** **4 marks****28 (12 marks)****(a)** $\text{Var}(\hat{y}) = \text{Var}(\bar{y}) + (x_0 - \bar{x})^2 \text{Var}(b)$ **M1 A1**

$$= \frac{\sigma^2}{n} + \frac{\sigma^2(x_0 - \bar{x})^2}{S_{xx}}$$

A1**Answer: $\sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$** **3 marks****(b)** $x_0 = 5: \sqrt{(0.1440)} = 0.379$ **M1 A1** $x_0 = 9: 0.651$ **A1** $x_0 = 20: 2.018$ **A1****Answer: 0.379, 0.651, 2.018****4 marks****(c)** The standard error is smallest at $\bar{x}$ and increases with the distance from $\bar{x}$ **E1**At $x = 20$ it is 5.3 times that at $\bar{x}$, so predictions far outside the data have much greater uncertainty, even before considering whether the model still holds**E1****Answer: Uncertainty grows with distance****2 marks**

$$\text{(d)} \quad \frac{1}{n} + \frac{(x_0 - 5)^2}{82.5} = 2 \times \frac{1}{10}$$

M1 $(x_0 - 5)^2 = 8.25$, so $x_0 = 5 \pm 2.872$, i.e. 2.13 or 7.87**M1 A1****Answer: $x_0 = 2.13$ or 7.87** **3 marks**

29 (11 marks)**(a)** Differences: 2.7, 1.9, 2.1, 2.8, 1.8, 2.8, 2.9, 1.8

M1 A1

Mean = 2.350, standard deviation = 0.493

M1 A1

Standard error = $\frac{0.493}{\sqrt{8}} = 0.1742$, $t_7 = 2.365$

M1

Interval = (1.938, 2.762)

A1

Answer: (1.94, 2.76)

6 marks

(b) $\sqrt{\left(\frac{3.696^2}{8 + 3.425^2/8}\right)} = 1.781$

M1 A1

Answer: 1.781

2 marks

(c) The paired standard error 0.174 is far smaller than 1.781

B1

The same random numbers make the two systems experience the same random variation, so the differences remove the shared variability and are much less variable than the individual times

E1 E1

Answer: Paired SE 0.174 against 1.781

3 marks

30 (12 marks)**(a)** $D = \frac{29 \times 7.1}{4.2} = 49.02$

M1 A1

Answer: 49.02

2 marks

(b) $49.02 > 42.56$

M1

There is significant evidence of over-dispersion, so the Poisson model is not appropriate

A1 E1

Answer: Reject Poisson: over-dispersed

3 marks

(c) $\mu = 4.2$ and $4.2 + \frac{4.2^2}{k} = 7.1$

M1

 $\frac{4.2^2}{k} = 2.90$ so $k = 6.083$

M1 A1

Answer: k = 6.08

3 marks

(d) Poisson: $e^{-4.2} = 0.0150$; negative binomial: $(6.08/10.28)^{6.08} = 0.0410$

M1 A1

The negative binomial model predicts 2.7 times as many fault-free days, as well as more days with many faults, so a Poisson model would misjudge both extremes and planning based on it would be unreliable

A1 E1

Answer: 0.015 and 0.041

4 marks