

Statistical Modelling

- 1 State what is meant by a statistical model, and give one reason why such a model is used. [2]
- 2 Give two reasons why a simple model may be preferred to a more complicated one. [2]
- 3 The number of emails received by a company in each ten-minute period is to be modelled by a Poisson distribution. State two assumptions needed for this model to be suitable. [2]
- 4 A sample of 20 items is taken from a production line and the number of defective items is modelled by a binomial distribution. State two conditions needed for this to be a suitable model. [2]
- 5 A fair six-sided die is rolled and the score X is modelled by a discrete uniform distribution on 1, 2, 3, 4, 5, 6. Write down $P(X = 4)$, $E(X)$ and $\text{Var}(X)$. [3]
- 6 For each situation, suggest a suitable distribution: (i) the number of heads in 15 tosses of a fair coin, (ii) the number of faults in a 100 m length of cable, (iii) the time between two successive arrivals at a bus stop. [3]
- 7 The number of faulty bulbs in a box of 12 is modelled by $B(12, 0.3)$. Find the probability that a box has at least 4 faulty bulbs. [3]
- 8 The number of calls to a helpline in a minute is modelled by a Poisson distribution with mean 2.4. Find the probability that in a given minute there are (i) no calls, (ii) at least two calls. [4]
- 9 The heights of adult men in a town are modelled by a normal distribution with mean 172 cm and standard deviation 8 cm. Find the proportion of men taller than 185 cm. [3]
- 10 A binomial model $B(5, p)$ is proposed for the number of successes in each of 80 groups of 5 trials. The results are shown in the table.
Estimate p from the data, and use it to find the expected frequency of groups with exactly 2 successes.

Successes	0	1	2	3	4	5
Frequency	3	14	27	22	11	3
- 11 The table shows the number of faults found in each of 100 rolls of fabric. A Poisson model is proposed. Estimate the mean of the model from the data, and calculate the expected number of rolls with no faults under this model.

Faults	0	1	2	3	4	5
Frequency	18	31	26	15	7	3
- 12 The time T minutes between successive customers entering a shop is modelled by an exponential distribution with rate 0.4 per minute. Find $P(T > 3)$ and $P(2 < T < 5)$. [4]
- 13 A lift waits at the ground floor for a time T seconds which is modelled by a continuous uniform distribution on $[0, 30]$. Find $P(T > 22)$, $E(T)$ and the standard deviation of T . [4]

- 14** The table shows the number of arrivals in each of 100 time intervals. The mean of the data was calculated as 1.6 and a Poisson model with $\lambda = 1.6$ is proposed. Calculate the expected frequencies, and test at the 5% significance level whether the Poisson model is a good fit.

Number	0	1	2	3	4 or more
Observed	22	31	24	14	9

[6]

- 15** A bag holds 3 red and 7 blue counters. Four counters are taken from the bag without replacement, and a student models the number of red counters by $B(4, 0.3)$. Explain why the model is not exactly correct, and find the true probability of exactly 2 red counters and the probability given by the model.

[4]

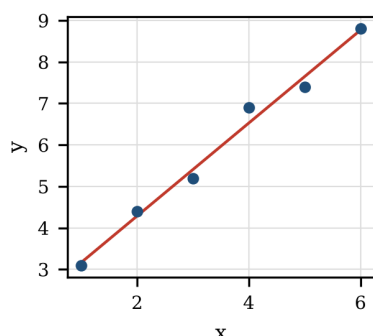
- 16** The number of breakdowns of a machine per week over 40 weeks has mean 3.1 and variance 6.8. A Poisson model is proposed. Comment on the suitability of the model, and suggest a reason why it may fail.

[4]

- 17** A student fits a Poisson distribution with mean 1.7 to 100 observations. For the final class “3 or more” the student calculates the expected frequency as $100 \times P(X = 3) = 14.96$. Identify the student’s error and find the correct expected frequency.

[4]

- 18** Data on the temperature x ($^{\circ}\text{C}$) in a greenhouse and the growth y (cm) of seedlings are shown: $x = 1, 2, 3, 4, 5, 6$ and $y = 3.1, 4.4, 5.2, 6.9, 7.4, 8.8$. A linear model $y = a + bx$ is proposed. Find the least squares values of a and b , and use the model to predict the growth when $x = 4.5$.



[6]

- 19** A journey consists of a bus ride followed by a walk. The bus time A is modelled by $N(12, 2^2)$ minutes and the walking time B by $N(8, 1.5^2)$ minutes, independently. Find the probability that the total journey takes more than 23 minutes.

[5]

- 20** The number of calls per hour to two independent switchboards are modelled by $Po(3.2)$ and $Po(2.1)$. Find the probability that the total number of calls in an hour is at least 8, and state why independence is needed for this method.

[5]

- 21** The table shows the number of goals scored in each of 100 football matches. A Poisson model is proposed. In the calculation of the mean, take “5 or more” as 5.

Goals	0	1	2	3	4	5 or more
Matches	14	31	30	15	7	3

- (a) Estimate the mean of the Poisson model. [2]
 (b) Calculate the expected frequencies for the six classes. [2]
 (c) Test at the 5% significance level whether the Poisson model fits, combining classes where necessary. [5]

- 22** The times between arrivals at a clinic are modelled by an exponential distribution. For a sample of 40 gaps the mean was 2.5 minutes.
- Estimate the rate parameter, and find the probability that a gap is longer than 5 minutes. [3]
 - A patient has already waited 3 minutes since the last arrival. Find the probability that the gap lasts at least 2 minutes more, and comment on the result. [4]
 - Give one reason why this property may make the model unsuitable for arrivals at a clinic with appointments. [1]

- 23** A factory packs 200 items in each box, and each item is independently faulty with probability 0.015.
- Find, using a binomial model, the probability that a box contains at most 2 faulty items. [2]
 - Use a Poisson approximation to find the same probability. [2]
 - Find the percentage error in the approximation, and explain why a Poisson model is reasonable here. [3]

- 24** A student tests whether a Poisson distribution fits data in 6 classes, estimating the mean from the data. The student finds $\chi^2 = 9.6$ and writes: "There are 6 classes so $\nu = 6 - 1 = 5$. The critical value at 5% is 11.070. Since $9.6 < 11.070$ the Poisson model is a good fit."
- Identify the error in the student's working. [2]
 - State the correct critical value and the correct conclusion. [3]
 - Explain why a model that passes such a test is not necessarily the right model. [2]

- 25** A population y (in thousands) of bacteria is recorded at times x hours: the data are shown in the table. A model of the form $y = ab^x$ is proposed.

x	0	1	2	3	4	5
y	12	17	25	36	52	76

- By taking logarithms, show that $\ln y$ is a linear function of x , and find the least squares values of a and b . [5]
 - Use the model to predict the population after 8 hours, and give one reason why this prediction may be unreliable. [3]
- 26** The numbers of accidents per week at two junctions A and B are modelled by independent Poisson distributions with means 1.8 and 2.7.
- Find the probability that there are at least 6 accidents in total in one week. [3]
 - Find the probability that there are exactly 20 accidents in total over four weeks. [3]
 - State two reasons why the model may not hold in practice. [2]

- 27** A random sample $x_1, x_2, \dots, x_n$ is taken from a Poisson distribution with unknown mean λ .
- Write down the log-likelihood function and show that the maximum likelihood estimate of λ is the sample mean. [5]
 - Show that the estimator is unbiased, and find an expression for its variance. [4]
 - For $n = 50$ and $\sum x_i = 140$, find the estimate of λ and its estimated standard error. [2]

- 28** The number of items X bought by a customer who buys at least one item is modelled by a zero-truncated Poisson distribution, with $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!(1 - e^{-\lambda})}$ for $x = 1, 2, 3, \dots$

- Show that $E(X) = \frac{\lambda}{1 - e^{-\lambda}}$. [3]
- A sample of customers has mean 2.4 items. Show that the estimate of λ satisfies $\lambda = 2.4(1 - e^{-\lambda})$, and use the iteration $\lambda_{n+1} = 2.4(1 - e^{-\lambda_n})$ with $\lambda_0 = 2.4$ to find λ to 3 significant figures. [4]
- Using your estimate, find the probability that a customer buys exactly one item. [2]

- 29** Customers at a shop are of two types. A proportion 0.6 are occasional customers, whose number of purchases per month is modelled by $Po(2)$, and a proportion 0.4 are regular customers, whose number is modelled by $Po(6)$. The number X of purchases made by a customer chosen at random is modelled as a mixture of the two distributions.
- Find $E(X)$. [2]
 - Find $E(X^2)$ and $Var(X)$, and show that the variance is greater than the mean. [5]
 - Find the probability that a randomly chosen customer makes no purchases. [2]
 - Explain how this model helps when a single Poisson model is found to be over-dispersed. [2]

- 30** The table shows the times t , in minutes, between 100 successive customers. The sample mean was 2.1 minutes. An exponential model is proposed.

Time t	0-1	1-2	2-3	3-5	over 5
Frequency	38	25	14	15	8

- Write down the estimate of the rate parameter. [1]
- Calculate the expected frequency for each class. [4]
- Test at the 5% significance level whether the exponential model is a suitable fit. [5]
- Explain why the final class is "over 5" rather than a bounded interval, and why the rate parameter reduces the degrees of freedom. [2]

Simulation Techniques

- State what is meant by a simulation, and give one advantage of using one. [2]
- A simulation of a hospital waiting area is to be run on a computer instead of observing the real waiting area. Give two reasons why this may be preferred. [2]
- Random digits are used to simulate tosses of a fair coin, with digits 0 to 4 representing a head and 5 to 9 a tail. Use the digits 3 8 6 1 9 to simulate five tosses. [2]
- Random digits are used to simulate rolls of a fair die. Digits 1 to 6 give that score and the digits 0, 7, 8, 9 are ignored. Use the digits 4 0 7 2 9 6 1 5 to write down the scores of the first four rolls. [2]
- An event has probability 0.35. Random numbers from the interval (0, 1) are used to simulate it, with the event occurring if the number is less than 0.35. For the numbers 0.412, 0.847, 0.209, 0.356, 0.718, state for each number whether the event occurs. [2]
- Explain why a simulation should be run many times before conclusions are drawn. [2]
- A discrete random variable X has the distribution shown. Two-digit random numbers 00 to 99 are used to simulate values of X .
Allocate the random numbers to the values of X , and use the numbers 63 08 91 47 22 to simulate five values of X .

x	0	1	2	3
$P(X = x)$	0.1	0.4	0.3	0.2

 [3]
- A bag contains 5 red, 3 blue and 2 green counters. A counter is drawn, its colour noted and it is replaced. Describe how single random digits can simulate this, and use the digits 6 1 9 3 8 to simulate five draws. [3]

- 9 Random digits 1 to 6 are used to simulate the scores on two fair dice, with the digits 0, 7, 8 and 9 ignored, and digits taken in pairs. Use the digits 4 9 2 0 5 6 8 1 3 7 6 6 2 5 to simulate the totals of four throws of the pair of dice. [3]
- 10 The number of calls X in a minute is modelled by $Po(1.5)$. Two-digit random numbers are to be used to simulate X . Allocate the numbers 00 to 99 to the values $X = 0, 1, 2, 3, 4$ and “5 or more”, using cumulative probabilities rounded to two decimal places, and simulate X using the random numbers 14 67 97. [4]
- 11 Customers arrive at a single till. The time since the previous arrival is 1 minute for digits 0–3, 2 minutes for 4–6 and 3 minutes for 7–9. The service time is 2 minutes for digits 0–4 and 3 minutes for 5–9. The first gap is measured from time 0, when the till opens. Using the digits 1 8 4 6 for the gaps and 2 7 5 9 for the service times of four customers, find the time each customer waits to be served and the mean wait. [5]
- 12 A simulation of 50 runs of an experiment gives the event A in 14 runs. Use this to estimate $P(A)$. The theoretical probability is 0.3: comment on the difference. [3]
- 13 Values from the standard normal distribution are used to simulate IQ scores which are modelled by $N(100, 15^2)$. Use the values $-0.52, 1.31, 0.07$ to simulate three IQ scores. [3]
- 14 Random digits are used to simulate the number of rolls of a fair die needed to get the first six. Digits 1 to 6 are scores and 0, 7, 8, 9 are ignored. Use the digits 4 1 9 6 3 0 6 6 2 5 1 1 3 6 to simulate three such experiments, and find the mean number of rolls. [4]
- 15 A shop sells a product. The daily demand is 0 for digits 0–1, 1 for digits 2–5, 2 for digits 6–8 and 3 for digit 9. The shop begins each day with a stock of 2 items. Use the digits 7 9 3 0 8 5 to simulate demand on six days, and find the total number of customers who cannot be supplied. [5]
- 16 A population has values 1, 2, 3 with probabilities 0.5, 0.3, 0.2. Digits 0–4 are used for the value 1, 5–7 for 2 and 8–9 for 3. Use the digits 2 8 5 9 0 6 4 7 1 to simulate three samples of size 3, and find the mean of each sample. [4]
- 17 Events A , B and C have probabilities 0.25, 0.35 and 0.40. A student allocates the two-digit numbers as A : 00–25, B : 26–60 and C : 61–99. Explain the error, and give a correct allocation. [4]
- 18 A fair coin is tossed. If it shows heads a fair six-sided die is rolled, and if it shows tails a fair four-sided die is rolled. One digit is used for the coin (0–4 heads, 5–9 tails) and then digits are used for the die (1–6 for the six-sided die, 1–4 for the four-sided die), ignoring digits not allowed. Use the digits 3 8 1 5 9 4 6 2 7 2 1 4 to simulate three repetitions, stating the coin result and the score each time. [5]
- 19 A particle starts at 0 and each step moves it +1 if the digit is 0–4 and –1 if the digit is 5–9. Use the digits 2 7 4 9 0 6 1 8 3 5 to simulate ten steps, and state the final position and the greatest distance from 0 reached. [4]
- 20 A simulation of 200 runs gives an estimated probability 0.32. Find the standard error of this estimate and an approximate 95% confidence interval for the true probability. State the number of runs needed to halve the width of the interval. [6]

- 21** The time T hours between breakdowns of a machine is modelled by an exponential distribution with rate 0.5, with cumulative distribution function $F(t) = 1 - e^{-0.5t}$ for $t \geq 0$.
- Show that if U is a random number from the interval $(0, 1)$, then $T = -2 \ln(1 - U)$ simulates a value of T . [3]
 - Use the random numbers 0.37, 0.82, 0.05, 0.64 to simulate four values of T , and find their mean. [3]
 - Explain why the mean of a large number of simulated values should be close to 2. [1]
-
- 22** Customers arrive at a single server. The gap between arrivals is 2, 3 or 4 minutes for random numbers 00–29, 30–69 and 70–99 respectively, and the service time is 2, 3 or 4 minutes for random numbers 00–39, 40–79 and 80–99 respectively. The first customer arrives after the first gap, measured from time 0. Customers are served in order of arrival.
- | | | | |
|------------------|----------|----------|----------|
| Gap (min) | 2 | 3 | 4 |
| Numbers | 00–29 | 30–69 | 70–99 |
- Use the random numbers 12 55 81 34 07 90 for the gaps and 45 23 88 61 09 72 for the service times to simulate six customers, finding each customer's waiting time before service and the mean waiting time. [6]
 - Explain why this simulation of six customers is not enough to decide whether more servers are needed. [1]
-
- 23** A simulation estimates a probability p . A pilot run suggests that p is about 0.3. The estimate from n runs is to be within 0.02 of the true value with 95% confidence.
- Find the number of runs required using the pilot value. [3]
 - Find the number of runs needed if no pilot information is available, and explain why this value is safe to use. [2]
 - Explain why very high accuracy is expensive in a simulation. [2]
-
- 24** A student simulates an exponential time T with mean 4 minutes using the formula $T = \frac{\ln U}{0.25}$, where U is a random number from $(0, 1)$. For $U = 0.3$ the student obtains $T = -4.82$.
- Explain why the student's formula cannot be correct. [2]
 - Write down a correct formula, and use it to simulate T for $U = 0.3$ and for $U = 0.8$. [4]
 - Show that using $-4 \ln U$ and $-4 \ln(1 - U)$ are both valid. [2]
-
- 25** Cars pass a point at random with a mean rate of 3 per hour. The times between cars are modelled by an exponential distribution with rate 3.
- Write down a formula to simulate a time between cars (in hours) from a random number U on $(0, 1)$. [2]
 - Use the random numbers 0.62, 0.35, 0.8, 0.21, 0.55, 0.4 in order to simulate the arrival times of successive cars, and find the number of cars in the first hour. [4]
 - State the distribution of the number of cars in an hour, and why the simulation should agree with it. [1]
-
- 26** X and Y are independent and each is uniformly distributed on $(0, 1)$. A simulation is used to estimate $P(X + Y > 1.5)$.
- Use the six pairs $(X, Y) = (0.82, 0.77), (0.15, 0.31), (0.64, 0.95), (0.58, 0.49), (0.91, 0.7), (0.27, 0.88)$ to estimate the probability. [3]
 - Find the exact value of $P(X + Y > 1.5)$ using the unit square. [3]
 - Comment on the difference between your simulated estimate and the exact value. [2]

- 27** A continuous random variable X has probability density function $f(x) = 2x$ for $0 \leq x \leq 1$, and $f(x) = 0$ otherwise.
- Find the cumulative distribution function $F(x)$ for $0 \leq x \leq 1$. [2]
 - Show that if U is uniform on $(0, 1)$ then $X = \sqrt{U}$ has the distribution of X , and explain why the inverse of F gives a simulation method in general. [4]
 - Simulate four values of X using $U = 0.16, 0.49, 0.81, 0.04$ and compare their mean with $E(X)$. [5]
-
- 28** A random variable X has probability density function $f(x) = 6x(1 - x)$ for $0 \leq x \leq 1$. A rejection method generates pairs (U_1, U_2) of independent uniform numbers on $(0, 1)$, and accepts $X = U_1$ if $U_2 \leq \frac{f(U_1)}{1.5}$.
- Show that the maximum of $f(x)$ is 1.5, and that the acceptance probability for each pair is $\frac{2}{3}$. [4]
 - Apply the method to the pairs $(0.2, 0.7), (0.5, 0.95), (0.9, 0.5), (0.35, 0.8), (0.6, 0.98)$, stating which values of X are accepted. [4]
 - Find the expected number of pairs needed to obtain 100 accepted values, and explain why the rejection method is needed if the inverse of the cumulative distribution function cannot be written down. [3]
-
- 29** Three fair dice are rolled, and M is the largest of the three scores. A simulation is to estimate $E(M)$.
- Show that $P(M \leq m) = \frac{m^3}{216}$, and hence find $P(M = m)$. [4]
 - Find the exact value of $E(M)$ and $\text{Var}(M)$. [5]
 - The mean of n simulated values of M is used to estimate $E(M)$. Find the least n for which the standard error of the estimate is at most 0.05. [3]
-
- 30** A pilot simulation of a queueing system gives the event “the queue exceeds 5 customers” in 45 of 250 runs.
- Estimate the probability, and find an approximate 99% confidence interval for it. [5]
 - Find the number of runs needed to estimate the probability to within ± 0.01 with 99% confidence, using the pilot estimate. [3]
 - Compare the cost of this simulation with one that is accurate to within ± 0.02 , and suggest one way to reduce the error without increasing the number of runs. [3]

Random Number Methods

- A linear congruential generator is defined by $x_{n+1} = (ax_n + c) \bmod m$. State the name given to x_0 , and to m . [2]
- A generator uses $x_{n+1} = (5x_n + 3) \bmod 16$ with $x_0 = 7$. Find x_1, x_2 and x_3 . [3]
- A generator with modulus 16 produces the values 3, 12, 5, 14. Convert them into numbers between 0 and 1 by dividing by the modulus. [2]
- Explain what is meant by the period of a random number generator, and why a long period is wanted. [2]
- In the middle-square method, a four-digit number is squared, written as an eight-digit number (with leading zeros if needed), and the middle four digits are the next number. Find the number which follows 1234. [2]
- Explain why numbers produced by a computer algorithm for simulation are called pseudo-random. [2]

- 7** A generator uses $x_{n+1} = (3x_n + 1) \bmod 11$ with $x_0 = 2$. Write down the first 5 values after the seed and find the period. [4]
- 8** A generator uses $x_{n+1} = 7x_n \bmod 10$ with $x_0 = 3$. Find the first six values, state the period, and explain why the generator is unsuitable. [4]
- 9** A generator with modulus 16 gives the values 12, 15, 14, 9, 0. Use each value x to simulate a die score by taking the whole number part of $\frac{6x}{16}$ and adding 1. [3]
- 10** Uniform random numbers u from $(0, 1)$ are used to simulate a value from the continuous uniform distribution on $[5, 12]$ using $x = 5 + 7u$. Find the simulated values for $u = 0.12, 0.58, 0.93, 0.31$. [3]
- 11** Use the middle-square method with four-digit numbers, writing each square as an eight-digit number, to find the next three numbers after 2916. [4]
- 12** Ten random numbers from $(0, 1)$ are 0.61, 0.07, 0.88, 0.45, 0.3, 0.72, 0.19, 0.54, 0.96, 0.23. Find their mean and compare it with the value expected from a uniform distribution on $(0, 1)$. [3]
- 13** A generator uses $x_{n+1} = (4x_n + 1) \bmod 16$ with $x_0 = 0$. Find the first five values and explain why the generator is useless. [3]
- 14** A generator is $x_{n+1} = (21x_n + 3) \bmod 16$ with $x_0 = 0$. The conditions for a full period m are: c and m have no common factor, $a - 1$ is divisible by every prime factor of m , and $a - 1$ is divisible by 4 if m is. Show that the conditions hold here, and check by listing the sequence. [4]
- 15** A student uses $x_{n+1} = 12x_n \bmod 100$ with seed 0 and says: "The numbers look random because they change a lot." Explain two things wrong with this generator. [4]
- 16** A generator gives $u = 0.18, 0.74, 0.41, 0.93$. Use the inverse transform $T = -\frac{\ln(1-u)}{\lambda}$ to simulate values from an exponential distribution with $\lambda = 0.8$, and find their mean. [5]
- 17** A generator produced 100 digits and the number of times each digit occurred is shown. Test, at the 5% significance level, whether the digits are uniformly distributed.
- | Digit | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|-----------|---|---|----|---|---|----|---|----|---|----|
| Frequency | 4 | 8 | 11 | 7 | 9 | 12 | 6 | 13 | 9 | 11 |
- [5]
- 18** A runs test is used on 20 random numbers. A "run" is an unbroken sequence of numbers all above 0.5 or all below 0.5. The numbers are: 0.62, 0.48, 0.15, 0.81, 0.77, 0.3, 0.93, 0.26, 0.54, 0.09, 0.67, 0.88, 0.12, 0.45, 0.71, 0.36, 0.59, 0.04, 0.83, 0.22. Find the number of runs, and use the fact that, for n_1 numbers above and n_2 below, the number of runs has mean $\frac{2n_1n_2}{n} + 1$ and variance $\frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}$, to test at the 5% level using a normal approximation whether the numbers show a lack of randomness. [6]
- 19** A triangular distribution on $[0, 2]$ has probability density function $f(x) = x$ for $0 \leq x \leq 1$ and $f(x) = 2 - x$ for $1 < x \leq 2$. Show that its cumulative distribution function is $F(x) = \frac{x^2}{2}$ for $0 \leq x \leq 1$, and use the inverse of this part to simulate a value from $u = 0.18$. [5]

- 20** A binomial random variable $X \sim B(4, 0.35)$ is simulated by taking four random numbers from $(0, 1)$, counting each number below 0.35 as a success. Use the numbers 0.12, 0.71, 0.3, 0.55, 0.28, 0.93, 0.47, 0.19 to simulate two values of X , and state the mean of X . [4]
-
- 21** The Box–Muller method produces two independent standard normal values from independent uniform numbers u_1 and u_2 on $(0, 1)$ using $Z_1 = R \cos \theta$ and $Z_2 = R \sin \theta$, where $R = \sqrt{-2 \ln u_1}$ and $\theta = 2\pi u_2$.
- (a) Calculate Z_1 and Z_2 for $u_1 = 0.3$ and $u_2 = 0.15$. [5]
- (b) Use your results to simulate two values from a normal distribution with mean 50 and standard deviation 4. [2]
- (c) Show that $Z_1^2 + Z_2^2 = R^2 = -2 \ln u_1$, and state what this means for the size of the pair. [2]
-
- 22** The generator RANDU uses $x_{n+1} = 65539 x_n \bmod 2^{31}$.
- (a) Show that $x_{n+2} \equiv 6x_{n+1} - 9x_n \pmod{2^{31}}$. [4]
- (b) Hence explain why triples (u_n, u_{n+1}, u_{n+2}) from this generator are not independent. [3]
-
- 23** A generator is $x_{n+1} = (ax_n + 1) \bmod 32$ with $x_0 = 0$. It has full period 32 if $a - 1$ is divisible by 4.
- (a) Show that the conditions for full period hold for $a = 5$. [2]
- (b) Find the first eight values for $a = 5$, and the period for $a = 5$ and for $a = 3$. [5]
- (c) Give a reason why full period alone does not make the generator good. [1]
-
- 24** A student generates 40 numbers from $x_{n+1} = (21x_n + 3) \bmod 16$ with $x_0 = 0$ and finds that the digits of the numbers have a mean of 7.4, close to 7.5. The student concludes: “The generator is excellent for simulations needing millions of numbers.”
- (a) Find x_{16} and x_{17} , and state what they show. [3]
- (b) Explain why the student’s conclusion is wrong. [3]
- (c) Suggest how the generator could be improved. [1]
-
- 25** A Poisson random variable X with mean 2 is to be simulated from a uniform random number u using the inverse transform method: X is the smallest k with $P(X \leq k) \geq u$.
- (a) Find $P(X \leq k)$ for $k = 0, 1, 2, 3, 4, 5$. [3]
- (b) Simulate values of X for $u = 0.68, 0.13, 0.95$. [3]
- (c) Explain why this method gives each value of k with probability $P(X = k)$. [1]
-
- 26** A value from an approximately standard normal distribution can be simulated by adding twelve independent uniform numbers on $(0, 1)$ and subtracting 6.
- (a) Show that the result has mean 0 and variance 1. [3]
- (b) Simulate a value using 0.62, 0.35, 0.81, 0.07, 0.49, 0.93, 0.28, 0.55, 0.71, 0.14, 0.66, 0.39. [2]
- (c) State the theorem that justifies using the normal distribution, and one weakness of this method. [2]
-
- 27** A generator is $x_{n+1} = (ax_n + c) \bmod m$, where a, c, m and x_0 are integers with $0 \leq x_0 < m$.
- (a) Explain why, once a value has occurred twice in the sequence, the sequence must repeat from then on. [3]
- (b) Show that the sequence must repeat after at most m terms, and so the period is at most m . [3]
- (c) Show that if $c = 0$ and $x_0 = 0$ the period is 1, and that if $c = 0$ then $x_n = a^n x_0 \bmod m$. [4]
-

- 28** The Weibull distribution with parameters k and θ has cumulative distribution function $F(x) = 1 - e^{-(x/\theta)^k}$ for $x \geq 0$.
- (a) Show that $X = \theta(-\ln(1 - U))^{1/k}$ simulates a value from this distribution when U is uniform on $(0, 1)$. [4]
- (b) For $k = 2$ and $\theta = 3$ simulate values using $U = 0.2, 0.5, 0.9$. [3]
- (c) Use $U = 0.5$ to find the median of the distribution, and check that F gives $\frac{1}{2}$ there. [3]
-
- 29** A generator $x_{n+1} = (5x_n + 3) \bmod 16$ is used, and consecutive pairs (x_n, x_{n+1}) are plotted as points in the plane.
- (a) Show that every point lies on one of the lines $y = 5x + 3 - 16k$ for an integer k . [3]
- (b) List the values of k that occur over the full cycle of 16 values, and state how many different lines are used. [3]
- (c) Explain why this behaviour is a weakness, and how a large modulus reduces the problem. [3]
-
- 30** Two independent random variables U and W are such that U is uniformly distributed on $(0, 1)$ and W is any continuous random variable on $(0, 1)$. Let $V = (U + W) \bmod 1$, which is the fractional part of $U + W$.
- (a) Show that, for $0 \leq v \leq 1$ and a fixed value w of W , $P(V \leq v | W = w) = v$. [5]
- (b) Deduce that V is uniformly distributed on $(0, 1)$ whatever the distribution of W . [3]
- (c) Three generators give 0.73, 0.58 and 0.91. Find the combined value $(u_1 + u_2 + u_3) \bmod 1$, and give one advantage of combining generators. [3]

Monte Carlo Simulation

- 1** Explain what is meant by Monte Carlo simulation. [2]
-
- 2** In a Monte Carlo estimate of π , 1000 random points are placed in the unit square and 782 fall inside the quarter circle $x^2 + y^2 < 1$. Find the estimate of π . [2]
-
- 3** A curve lies inside a rectangle of area 60. Random points are placed in the rectangle and 37 out of 100 fall under the curve. Estimate the area under the curve. [2]
-
- 4** X and Y are independent and uniformly distributed on $(0, 1)$. State the probability that the point (X, Y) lies inside the quarter circle $x^2 + y^2 < 1$, and explain how this is used to estimate π . [2]
-
- 5** Estimate $\int_0^1 x^2 dx$ by evaluating x^2 at the random numbers 0.2, 0.6, 0.9 and finding the mean. [2]
-
- 6** State one way of making a Monte Carlo estimate more accurate, and explain why it works. [2]
-
- 7** Estimate $\int_0^2 x^2 dx$ using the random numbers 0.15, 0.45, 0.7, 0.95 from $(0, 1)$, by first converting each to a value of x in $(0, 2)$. Compare with the exact value. [5]
-
- 8** In 200 simulated throws of two dice, a total of at least 10 occurred 31 times. Estimate the probability, and compare with the exact value. [4]
-
- 9** Eight random points (x, y) in the unit square are (0.12, 0.64), (0.85, 0.71), (0.4, 0.33), (0.91, 0.52), (0.28, 0.95), (0.66, 0.2), (0.07, 0.08), (0.77, 0.58). Count how many lie inside the quarter circle $x^2 + y^2 < 1$ and estimate π . [3]

- 10** In a Monte Carlo estimate of π using $n = 1000$ points, the proportion inside the quarter circle has standard error $\sqrt{\frac{p(1-p)}{n}}$ with $p = \frac{\pi}{4}$. Find this standard error, and the standard error of the resulting estimate of π . [4]
- 11** The region under $y = x^2$ for $0 \leq x \leq 1$ is inside the unit square. Of 40 random points in the square, 14 lie below the curve. Estimate the area under the curve and compare with the exact value. [4]
- 12** Ten runs of a simulation give the waiting times (minutes): 12.4, 15.1, 13.8, 14.6, 12.9, 15.7, 13.3, 14.2, 13.9, 14.7. Find the mean and the standard error of the mean. [4]
- 13** Two methods are available for estimating $\int_0^1 f(x) dx$ by Monte Carlo: the hit-or-miss method and the average value method. Describe each method. [2]
- 14** Use the average value method with the random numbers 0.1, 0.35, 0.55, 0.8, 0.95 to estimate $\int_1^3 \ln x dx$, and compare your estimate with the exact value. [6]
- 15** The maximum of two independent random numbers uniform on $(0, 1)$ is M . Use the six pairs (0.31, 0.72), (0.88, 0.14), (0.52, 0.6), (0.09, 0.45), (0.77, 0.9), (0.63, 0.25) to estimate $E(M)$, and show that the exact value is $\frac{2}{3}$. [5]
- 16** A Monte Carlo estimate of π uses 2000 random points in the unit square, of which 1573 are inside the quarter circle. Find the estimate of π and an approximate 95% confidence interval for π , and state whether the interval contains the true value. [6]
- 17** A student estimates $\int_0^3 e^x dx$ by evaluating e^x at $x = 3u$ for the random numbers $u = 0.2, 0.5, 0.8$ and finding the mean of the three values, 5.78. Identify the error and correct the estimate. [4]
- 18** Random points (x, y, z) in the unit cube are (0.42, 0.31, 0.77), (0.85, 0.52, 0.19), (0.16, 0.64, 0.38), (0.93, 0.71, 0.45), (0.27, 0.08, 0.69), (0.58, 0.83, 0.22), (0.34, 0.47, 0.51), (0.76, 0.9, 0.13). Use the number of points with $x^2 + y^2 + z^2 < 1$ to estimate the volume of a sphere of radius 1, and compare with the exact value. [6]
- 19** A system works if component A works and at least one of components B and C works. The components work independently with probabilities 0.9, 0.8 and 0.7. Find the exact probability that the system works, and use the random number triples (0.55, 0.12, 0.4), (0.93, 0.35, 0.66), (0.2, 0.81, 0.9), (0.71, 0.47, 0.29) (each number below the probability means the component works) to estimate it by simulation. [6]
- 20** The average value method estimates $\int_0^1 x^2 dx$ using the mean of U^2 for n random numbers U . Show that the variance of U^2 is $\frac{4}{45}$ and find the standard error of the estimate when $n = 1000$. [5]
- 21** The integral $I = \int_0^1 x^2 dx = \frac{1}{3}$ is estimated using n random numbers by two methods. In the hit-or-miss method, random points in the unit square are used and the proportion under the curve $y = x^2$ estimates I . In the average value method, the mean of U^2 is used.
- (a) Find the variance of a single observation in the hit-or-miss method. [2]
 - (b) Find the variance of a single observation in the average value method. [2]
 - (c) Find how many times as many random points are needed by the hit-or-miss method to achieve the same accuracy, and comment. [3]

- 22** The integral $\int_0^1 e^x dx$ is estimated using the average value method. An antithetic estimator uses $Y = \frac{e^U + e^{1-U}}{2}$ for each random number U , where U is uniform on $(0, 1)$.
- Show that $E(Y) = e - 1$, and explain why Y is an unbiased estimator of the integral. [3]
 - Given that $\text{Var}(e^U) = 0.2420$ and $\text{Cov}(e^U, e^{1-U}) = -0.2342$, find $\text{Var}(Y)$. [3]
 - Compare the variance of Y with the variance of e^U , and explain why the antithetic method works here. [2]
-
- 23** In Buffon's needle experiment, a needle of length l is dropped on a floor ruled with parallel lines a distance d apart, where $l \leq d$. The probability that the needle crosses a line is $\frac{2l}{\pi d}$.
- A needle of length 1 is dropped 400 times on lines 1 unit apart and crosses a line 255 times. Estimate π . [3]
 - Find the standard error of the proportion, and hence a 95% confidence interval for π . [4]
 - State whether the interval contains π , and explain why the interval for π is not symmetric about the estimate. [2]
-
- 24** A student runs a Monte Carlo estimate of π using 500 points and obtains 3.18. The student writes: "The simulation gives 3.18, so the value 3.1416 is wrong and π is about 3.18."
- Find the standard error of the student's estimate of π . [3]
 - Show that the true value of π is consistent with the student's result. [3]
 - Explain the student's error, and suggest how to get a result nearer to the true value. [2]
-
- 25** X is exponentially distributed with rate 1. A Monte Carlo method estimates $E(\sqrt{X})$.
- State a formula for simulating X from a uniform random number u , and use it with $u = 0.72, 0.15, 0.4, 0.91$ to simulate four values of X . [3]
 - Hence estimate $E(\sqrt{X})$. [2]
 - The exact value of $E(\sqrt{X})$ is $\Gamma(\frac{3}{2}) = \frac{\sqrt{\pi}}{2}$. Find it, and comment on the simulation. [2]
-
- 26** A region lies within a rectangle of area 20 square units. In a Monte Carlo simulation, 10 000 random points are placed in the rectangle and 3120 fall inside the region.
- Estimate the area of the region. [2]
 - Find a 95% confidence interval for the area. [4]
 - Explain how the accuracy could be improved without changing the number of points. [1]
-
- 27** A region of area A lies inside a rectangle of area R . In the hit-or-miss method, n independent random points are placed uniformly in the rectangle and H is the number which fall inside the region. The estimate of A is $\hat{A} = \frac{RH}{n}$.
- Show that $\hat{A}$ is an unbiased estimator of A . [3]
 - Show that $\text{Var}(\hat{A}) = \frac{A(R - A)}{n}$. [4]
 - Find the value of A for which the variance is greatest for a given R , and hence show that a smaller rectangle reduces the variance when A is fixed. [4]

28

A function f satisfies $0 \leq f(x) \leq 1$ for $0 \leq x \leq 1$, and $I = \int_0^1 f(x) \, dx$. The hit-or-miss estimate uses the proportion of random points in the unit square which lie below the curve. The average value estimate uses the mean of $f(U)$ for random numbers U .

- (a) Show that the variance of one observation is $I(1 - I)$ for hit-or-miss and $\int_0^1 f(x)^2 \, dx - I^2$ for average value. [3]
- (b) Show that the average value method never has the larger variance. [3]
- (c) For which functions are the two variances equal? [2]
- (d) Calculate the two variances for $f(x) = x$, and the number of hit-or-miss points needed for the same accuracy as 1000 average value points. [3]

29

A Monte Carlo simulation estimates the tail probability $p = P(Z > 3)$ for a standard normal variable Z , where $p = 0.00135$, by simulating n values of Z and recording the proportion exceeding 3.

- (a) Find the standard error of the estimate in terms of n , and the relative standard error (the standard error divided by p). [3]
- (b) Find the least n for which the relative standard error is at most 10%. [3]
- (c) Explain why simulating rare events with a plain Monte Carlo method is expensive, and suggest how a different approach could help. [3]
- (d) In importance sampling the values are drawn from a normal distribution with mean 3 and standard deviation 1. Show that, for such a value $y > 3$, the weight is $e^{-3y + 4.5}$. [3]

30

In a Monte Carlo estimate of π , n points are placed in the unit square and π is estimated as $4H/n$, where H is the number inside the quarter circle.

- (a) Show that the estimator is unbiased, and that its variance is $\frac{\pi(4 - \pi)}{n}$. [4]
- (b) Find the standard error when $n = 10^6$. [2]
- (c) Find the number of points needed to be 99% sure that the estimate is correct to within 0.001, and comment on the efficiency of the method compared with numerical integration. [5]

Interpreting Statistical Models

- 1 A regression line $y = 12 + 3.5x$ is fitted to data where x is the number of hours of revision and y is the test mark. Interpret the values 12 and 3.5 in context. [2]
- 2 Define a residual in a regression model. [2]
- 3 The model $y = 2 + 3x$ is used. At $x = 12$ the observed value is 41. Find the predicted value and the residual. [2]
- 4 A plot of residuals against x shows no pattern, with points scattered randomly above and below zero. State what this suggests about the model. [2]
- 5 Explain why it is unreliable to use a regression line to predict y for values of x far outside the range of the data. [2]
- 6 In a linear regression the coefficient of determination is $r^2 = 0.81$. Interpret this value. [2]
- 7 The line $y = -4.2 + 0.8x$ models daily ice cream sales y (hundreds) at a kiosk when the temperature is x °C. Interpret the gradient and the intercept, and comment on the intercept. [3]

- 8** A model $y = 1.2 + 2.1x$ is proposed for the data in the table. Calculate the residuals and state whether the model overestimates or underestimates at $x = 7$.

x	2	4	5	7	9
y	5.6	9.9	11.0	16.2	19.5

[4]

- 9** Two models are proposed for the data $(x, y) = (3, 8.4), (5, 12.1), (8, 18.0), (10, 22.3)$: Model A is $y = 0.5 + 2.1x$ and Model B is $y = 1 + 2x$. Calculate the residual sum of squares for each model and state which fits better.

[4]

- 10** The residuals for increasing values of x are $+2.1, +1.4, +0.3, -1.2, -1.9, -1.1, +0.4, +1.8$. Describe the pattern and explain what it suggests about a linear model.

[3]

- 11** A model $y = 20 + 1.5x$ is fitted using data with x between 10 and 50. Find the predictions when $x = 35$ and $x = 80$, and say which one is more reliable.

[4]

- 12** A Poisson model is fitted to data and a goodness of fit test gives a p -value of 0.62. Interpret this result in context.

[2]

- 13** Twenty-five runs of a simulation give a mean waiting time of 8.4 minutes with a sample standard deviation of 2.5 minutes. Find a 95% confidence interval for the true mean waiting time.

[4]

- 14** A model $y = a + bx$ is fitted by least squares to the data in the table.

x	1	2	3	4	5	6
y	2.3	3.1	4.9	5.2	6.8	7.1

(a) Find a and b .

[3]

(b) Calculate the residuals and show that they sum to zero, and comment on whether they suggest the model is appropriate.

[4]

- 15** A model for the population y of a colony is $\ln y = 0.5 + 0.2x$, where x is the time in days. Show that the model implies $y = Ab^x$, find A and b , and interpret b in context.

[4]

- 16** In a simulation of a call centre the mean waiting time was 6.2 minutes. When the call rate is increased by 10% and the simulation is repeated, the mean waiting time is 8.9 minutes. Find the percentage change in waiting time, and explain what this shows about the model.

[4]

- 17** A student says: "The correlation coefficient is $r = 0.8$, so 80% of the points lie close to the line, and the data are 80% explained by the model." Identify two errors in this statement.

[4]

- 18** A regression line is fitted to seven points (x, y) with $x = 1, 2, \dots, 7$ and $y = 3.0, 4.2, 5.1, 5.8, 7.2, 8.0, 14.5$.

(a) Find the gradient of the least squares line using all seven points.

[3]

(b) Find the gradient when the point $(7, 14.5)$ is omitted, and comment on the point.

[4]

- 19** A linear model (2 parameters) and a quadratic model (3 parameters) are each fitted to 12 data points. The residual sums of squares are 120 for the linear model and 112 for the quadratic model. Calculate the estimate of the error variance $\frac{RSS}{n - p}$ for each model, and state which model should be preferred, giving a reason.

[5]

- 20** A model produces 40 standardised residuals, of which 3 lie outside the range -2 to 2 . If the errors are normal the proportion outside is about 4.6%. Find the number expected outside this range, and comment on whether the normality assumption appears reasonable.

[4]

- 21** A company records the advertising spend x (in £ hundreds) and the sales y (in £ thousands) for eight months, as shown in the table.

x	10	15	20	25	30	35	40	45
y	14.2	17.8	24.1	26.0	33.5	35.2	43.0	45.1

- (a) Find the equation of the least squares regression line of y on x . [3]
 (b) Calculate r and r^2 , and interpret r^2 . [3]
 (c) Predict sales for a spend of 28 (hundred pounds), and explain why a prediction for a spend of 80 would be unreliable. [3]

- 22** The table shows the number of defects found in each of 120 components. A Poisson model and a geometric model (with $P(X = k) = p(1 - p)^k$, $k = 0, 1, 2, \dots$) are both proposed. Take “4 or more” as 4 when finding the sample mean, which is 1.750.

Number	0	1	2	3	4 or more
Frequency	22	34	31	18	15

- (a) Fit the Poisson model, find the expected frequencies and the value of χ^2 . [4]
 (b) Fit the geometric model, using $p = \frac{1}{1 + \text{mean}}$, and find the value of χ^2 . [3]
 (c) Using the 5% critical value for $\nu = 3$, state which model is acceptable and which is better. [2]

- 23** A regression of y on x gives residuals which are grouped by fitted value. The standard deviation of the residuals is 1.2 for fitted values below 20, 2.5 for fitted values from 20 to 40 and 4.9 for fitted values above 40.

- (a) Describe what these values show about the model. [2]
 (b) Explain why this matters for the use of the model. [2]
 (c) Suggest a transformation which might help, and explain why. [3]

- 24** A manager runs a simulation 10 times and records the mean daily cost (in £ hundreds) each time: 11.2, 13.5, 12.8, 10.9, 14.1, 12.2, 13.0, 11.7, 12.6, 13.4. The manager says: “The mean of these runs is 12.54, so the true mean cost is exactly 12.54.”

- (a) Find the sample mean and standard deviation. [3]
 (b) Find a 95% confidence interval for the true mean cost. [3]
 (c) Explain the error in the manager’s statement. [2]

- 25** A stock model gives the annual cost of ordering in batches of size q as $C(q) = \frac{DS}{q} + \frac{hq}{2}$, with demand $D = 1200$ per year, ordering cost $S = 50$ and holding cost $h = 2$ per item per year.

- (a) Find the batch size q which minimises $C(q)$, and the minimum cost. [3]
 (b) The demand is found to be 20% higher than assumed, at 1440. Find the new best batch size and the cost of using the old batch size with the new demand. [3]
 (c) Calculate the percentage change in the best batch size and the percentage excess cost of the old batch size, and comment on the sensitivity. [3]

- 26** A model was built from earlier data. It is tested on six new observations which were not used to fit the model. The table shows the observed and predicted values.

Month	1	2	3	4	5	6
Observed	34	41	28	52	47	39
Predicted	36	38	31	49	50	38

- (a) Find the error (predicted – observed) for each month, and the mean error. [3]
 (b) Find the mean absolute error and the root mean square error. [4]
 (c) Explain why testing on new data is better than judging a model by how well it fits the data used to build it. [1]

27

A least squares line $y = a + bx$ is fitted to n points (x_i, y_i) by minimising $S = \sum_{i=1}^n (y_i - a - bx_i)^2$. The residuals are $e_i = y_i - a - bx_i$.

- (a) By differentiating S with respect to a and b , show that $\sum_{i=1}^n e_i = 0$ and $\sum_{i=1}^n x_i e_i = 0$. [4]
 (b) Hence show that the fitted line passes through the point $(\bar{x}, \bar{y})$. [2]
 (c) Show that the residual sum of squares is $S_{yy} - \frac{S_{xy}^2}{S_{xx}}$, and hence that $r^2 = 1 - \frac{RSS}{S_{yy}}$. [4]

28

A regression model $y = \alpha + \beta x + \varepsilon$, with independent errors of variance σ^2 , is fitted to n observations. The fitted line is $\hat{y} = \bar{y} + b(x - \bar{x})$, where $\text{Var}(\bar{y}) = \frac{\sigma^2}{n}$, $\text{Var}(b) = \frac{\sigma^2}{S_{xx}}$, and $\bar{y}$ and b are independent.

- (a) Show that the variance of the predicted mean response at $x = x_0$ is $\sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$. [3]
 (b) For $n = 10$, $\bar{x} = 5$, $S_{xx} = 82.5$ and $\sigma = 1.2$, find the standard error of the predicted mean response at $x_0 = 5$, $x_0 = 9$ and $x_0 = 20$. [4]
 (c) Explain what this shows about prediction far from $\bar{x}$. [2]
 (d) Find the value of x_0 for which the variance is twice its minimum value. [3]

29

Two designs A and B of a system are compared by simulation. Each of eight runs uses the same sequence of random numbers for both designs, and the mean times (in minutes) are shown. The aim is to estimate the true mean difference $\mu_A - \mu_B$.

Run	1	2	3	4	5	6	7	8
System A	34.2	36.8	29.5	41.0	33.7	38.4	35.9	31.6
System B	1.5	34.9	27.4	38.2	31.9	35.6	33.0	29.8

- (a) Find the differences, their mean and standard deviation, and a 95% confidence interval for the mean difference using the paired differences. [6]
 (b) Find the standard errors of the difference if the runs were treated as independent samples, using $\sqrt{\frac{s_A^2}{8} + \frac{s_B^2}{8}}$, where $s_A = 3.696$ and $s_B = 3.425$. [2]
 (c) Compare the two standard errors, and explain why using the same random numbers for both designs gives a more precise comparison. [3]

30 The number of faults per day in a factory over 30 days has sample mean 4.2 and sample variance 7.1. A manager proposes a Poisson model.

- (a) Calculate the dispersion statistic $\frac{(n-1)s^2}{\bar{x}}$, which is approximately χ^2 with $n-1$ degrees of freedom if the data are Poisson. [2]
- (b) Test at the 5% significance level whether the data are consistent with a Poisson distribution, using the upper-tail critical value of χ^2 with 29 degrees of freedom, which is 42.56. [3]
- (c) A negative binomial model with mean μ and variance $\mu + \frac{\mu^2}{k}$ is proposed. Estimate k by equating the model mean and variance to the sample values. [3]
- (d) The probability of a day with no faults is $\frac{k}{k+\mu}$ raised to the power k under the negative binomial model, and $e^{-\mu}$ under the Poisson model. Calculate both, and explain why the choice of model matters for planning. [4]