Normal Distribution — Shadow Questions

1. Resting Heart Rate

The resting heart rate for adults is normally distributed with mean 70 bpm and standard deviation 12 bpm. Find the probability that a randomly chosen adult has:

(a) a heart rate under 95 bpm.

(b) an abnormal heart rate outside the range 45–95 bpm.

2. Examination Grades

Some examinations are graded in terms of the mean μ and standard deviation σ of the marks. The lowest marks for grades A–U in a geography examination are:

Grade	A	B	C	D	U
Minimum mark	$\mu + 2\sigma$	$\mu + \sigma$	$\mu - \sigma$	$\mu - 2\sigma$	0

(a) Estimate the proportion of candidates graded D. State any assumptions.

(b) The exam was taken by 600 students. Scholarships of £3500 were awarded to each student who achieved grade A. Find the expected total cost.

3. Train Arrival Times

The arrival times of a weekday train are normally distributed with mean 08:52 and standard deviation 5 minutes. In one year, how often would you expect the train to arrive after 09:00?

4. Connecting Rods

The lengths of 250 mm connecting rods from supplier X have mean 250 mm and standard deviation 0.5 mm.

(a) A sample of twelve rods from another supplier have lengths (mm):
249.6, 250.3, 249.9, 250.5, 249.7, 250.0, 249.8, 250.1, 250.2, 250.4, 250.3, 249.8
Compare these rods with those from supplier X. Would you switch supplier?

(b) Any rod with length outside $249 \leq \ell \leq 251$ mm is unusable. In a batch of 3200 rods from supplier X, how many would you expect to be unusable?

5. BMI Classification

Adults with BMI over 25 are classified as overweight; over 30 as obese. In the UK, 62% of adults are overweight and 25% are obese.

(a) A large sample suggests BMIs follow $N(26.8, 5.1^2)$. Check whether this is consistent with the given percentages. Comment on your results and assumptions.

(b) A sample of 10 male patients in a clinic have BMIs:
33, 38, 24, 21, 29, 22, 35, 31, 26, 37
Compare these patients with the general UK adult population.

Confidence Intervals — Shadow Questions

1. Standard Error

A random sample of size n is taken from $N(\mu, \sigma^2)$. Which of the following is the standard error of the mean?

$$\frac{\sigma}{\sqrt{n}}, \qquad \frac{\sigma^2}{n}, \qquad \sqrt{\frac{\sigma}{n}}$$

2. Lettuce Weights

Weights of lettuces are normally distributed with mean 610 g and standard deviation 22 g.

(a) Find the probability that a randomly chosen lettuce weighs between 600 g and 620 g.

(b) Find the probability that the mean weight of four lettuces lies between 600 g and 620 g.

3. Radon Levels

Radon levels (Bq m^{-3}) in six houses are:

$$17.9,\ 19.4,\ 20.8,\ 21.1,\ 18.7,\ 20.2$$

Assuming radon levels follow $N(\mu, 1.4^2)$, construct a 98% confidence interval for μ.

4. Orange Weights

A supplier claims the mean weight of oranges is 205 g. A sample of 12 oranges has mean 198 g and known standard deviation 9 g. Find a 99% confidence interval for the mean and comment on the supplier's claim.

5. IQ Scores

IQ scores in a town follow $N(\mu, 20^2)$. The average IQ of eight married couples is:

$$94,\ 118,\ 129,\ 103,\ 134,\ 92,\ 99,\ 95$$

(a) Show how a student might estimate the distribution as $N(108, 20^2)$.

(b) What assumptions has the student made?

(c) Comment on whether these assumptions are reasonable.

Mark Scheme — Normal Distribution and Confidence Intervals

Normal Distribution — Solutions

Question 1 — Resting Heart Rate

The resting heart rate $X \sim N(70, 12^2)$.

(a) Probability $X < 95$ Standardise:

$$Z = \frac{X-\mu}{\sigma} = \frac{95-70}{12} = \frac{25}{12} \approx 2.08.$$

So

$$P(X < 95) = P(Z < 2.08) \approx 0.981.$$

(b) Probability outside 45–95 We want

$$P(X < 45) + P(X > 95).$$

Standardise both bounds.
Lower bound:

$$Z = \frac{45-70}{12} = \frac{-25}{12} \approx -2.08 \Rightarrow P(X < 45) \approx P(Z < -2.08) \approx 0.019.$$

Upper bound:

$$P(X > 95) = 1 - P(X < 95) \approx 1 - 0.981 = 0.019.$$

So

$$P(\text{outside } 45\text{–}95) \approx 0.019 + 0.019 = 0.038.$$

Question 2 — Examination Grades

Assume marks are normally distributed with mean μ and standard deviation σ.

(a) Proportion graded D Grade D corresponds to marks between $\mu - 2\sigma$ and $\mu - \sigma$. Standardising:

$$P(\mu - 2\sigma \le X < \mu - \sigma) = P(-2 \le Z < -1).$$

From standard normal tables:

$$P(Z < -1) \approx 0.159, \quad P(Z < -2) \approx 0.023.$$

So

$$P(-2 \leq Z < -1) \approx 0.159 - 0.023 = 0.136.$$

Assumption: marks are exactly normally distributed and the grade boundaries are set at integer multiples of σ.

(b) Expected bursary cost Grade A corresponds to $X \geq \mu + 2\sigma$:

$$P(X \geq \mu + 2\sigma) = P(Z \geq 2) = 1 - P(Z < 2) \approx 1 - 0.977 = 0.023.$$

Number of A grades out of 600:

$$600 \times 0.023 \approx 13.8 \approx 14 \text{ students (expected).}$$

Each receives £3500, so expected cost:

$$14 \times 3500 = 49{,}000 \quad (\text{or } 600 \times 0.023 \times 3500 \approx 48{,}300).$$

Question 3 — Train Arrival Times

Let T be arrival time in minutes after 08:00. Mean is 08:52 $\Rightarrow \mu = 52$ minutes, $\sigma = 5$ minutes. We want $P(T > 60)$ (after 09:00).

$$Z = \frac{60 - 52}{5} = \frac{8}{5} = 1.6.$$

$$P(T > 60) = P(Z > 1.6) = 1 - P(Z < 1.6) \approx 1 - 0.945 = 0.055.$$

In one year (about 365 weekdays if daily, or ≈ 260 if weekdays only): - If daily: $365 \times 0.055 \approx 20$ times. - If weekdays only: $260 \times 0.055 \approx 14$ times.

Question 4 — Connecting Rods

Supplier X: $L_X \sim N(250, 0.5^2)$.

(a) Compare suppliers Other supplier sample (12 rods):

Data: $249.6, 250.3, 249.9, 250.5, 249.7, 250.0, 249.8, 250.1, 250.2, 250.4, 250.3, 249.8.$

Compute sample mean:

$$\bar{x} = \frac{\sum x_i}{12}.$$

Sum:

$$249.6+250.3+249.9+250.5+249.7+250.0+249.8+250.1+250.2+250.4+250.3+249.8 = 3000.6.$$

So

$$\bar{x} = \frac{3000.6}{12} = 250.05 \text{ mm (approx)}.$$

Sample standard deviation (using calculator or formula) is close to 0.3–0.4 mm (you can show):

$$s \approx 0.3\text{–}0.4 \text{ mm}.$$

Comparison: - Mean is very close to 250 mm (similar to supplier X). - Standard deviation appears slightly smaller than 0.5 mm, suggesting slightly more consistency.

Conclusion: On this evidence, the new supplier is at least as good, possibly better in consistency. A switch could be justified, but the sample is small.

(b) Unusable rods from supplier X Unusable if $L_X < 249$ or $L_X > 251$.

Standardise:

Lower bound:

$$Z = \frac{249 - 250}{0.5} = -2.$$

Upper bound:

$$Z = \frac{251 - 250}{0.5} = 2.$$

So

$$P(249 \leq L_X \leq 251) = P(-2 \leq Z \leq 2) \approx 0.954.$$

Thus

$$P(\text{unusable}) = 1 - 0.954 = 0.046.$$

Expected unusable rods in 3200:

$$3200 \times 0.046 \approx 147.2 \approx 147 \text{ rods}.$$

Question 5 — BMI Classification

Given: overweight > 25, obese > 30. UK: 62% overweight, 25% obese. Sample suggests $X \sim N(26.8, 5.1^2)$.

(a) Check consistency Overweight: $P(X > 25)$.

$$Z = \frac{25 - 26.8}{5.1} \approx \frac{-1.8}{5.1} \approx -0.35.$$

$$P(X > 25) = 1 - P(Z < -0.35) = 1 - (1 - P(Z < 0.35)) = P(Z < 0.35) \approx 0.636.$$

So model predicts about 63.6% overweight vs 62% given — reasonably close.
Obese: $P(X > 30)$.

$$Z = \frac{30 - 26.8}{5.1} \approx \frac{3.2}{5.1} \approx 0.63.$$

$$P(X > 30) = 1 - P(Z < 0.63) \approx 1 - 0.735 = 0.265.$$

So model predicts about 26.5% obese vs 25% given — again close.

Comment: The normal model with mean 26.8 and s.d. 5.1 is broadly consistent with the quoted percentages, assuming a normal distribution and a representative sample.

(b) Clinic sample comparison Clinic BMIs:

$$33, 38, 24, 21, 29, 22, 35, 31, 26, 37.$$

Compute mean:

$$\sum = 33 + 38 + 24 + 21 + 29 + 22 + 35 + 31 + 26 + 37 = 296.$$

$$\bar{x} = \frac{296}{10} = 29.6.$$

This is higher than the population mean 26.8, suggesting this clinic sample is heavier on average.

Proportion obese in clinic (BMI > 30): Values > 30: $33, 38, 35, 31, 37 \to 5$ out of 10 $\to$ 50% obese.

Compare with UK 25%: clinic has a much higher proportion of obese adults, consistent with a diabetes clinic population.

Question 6 — Soil pH

Field 1: $X \sim N(7.0, 0.3^2)$. Field 2 sample: $6.9, 6.6, 7.0, 6.8, 6.8, 6.7, 7.0, 6.8, 7.1, 6.7$.

(a) Compare suitability Compute mean for field 2:

$$\sum = 6.9 + 6.6 + 7.0 + 6.8 + 6.8 + 6.7 + 7.0 + 6.8 + 7.1 + 6.7 = 68.4.$$

$$\bar{x} = \frac{68.4}{10} = 6.84.$$

Field 1 mean: 7.0 (exactly ideal). Field 2 mean: 6.84 (slightly below ideal).

Standard deviation (field 2) is smaller than 0.3 (you can compute; it's around 0.15–0.2), so field 2 is more consistent but slightly more acidic on average.

Contextual comparison: - Field 1: mean exactly ideal, but more spread. - Field 2: slightly too acidic on average, but more uniform.

(b) Percentage too acidic in field 1 Too acidic if pH < 6.5.

$$Z = \frac{6.5 - 7.0}{0.3} = \frac{-0.5}{0.3} \approx -1.67.$$

$$P(X < 6.5) = P(Z < -1.67) \approx 0.048.$$

So about 4.8% of the field is expected to be too acidic.

Assumption: pH values are normally distributed and the mean and s.d. are accurate for the whole field.

Confidence Intervals — Solutions

Question 1 — Standard Error

For a sample of size n from $N(\mu, \sigma^2)$, the standard error of the mean is:

$$\text{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

So the correct expression is $\dfrac{\sigma}{\sqrt{n}}$.

Question 2 — Lettuce Weights

Weights: $X \sim N(610, 22^2)$.

(a) Single lettuce between 600 g and 620 g Standardise both bounds.

Lower:

$$Z_1 = \frac{600 - 610}{22} = \frac{-10}{22} \approx -0.455.$$

Upper:

$$Z_2 = \frac{620 - 610}{22} = \frac{10}{22} \approx 0.455.$$

So

$$P(600 < X < 620) = P(-0.455 < Z < 0.455).$$

From tables, $P(Z < 0.455) \approx 0.675$, so

$$P(-0.455 < Z < 0.455) \approx 0.675 - (1 - 0.675) = 0.675 - 0.325 = 0.350.$$

(b) Mean of 4 lettuces between 600 g and 620 g For sample mean $\bar{X}$ of size $n = 4$:

$$\bar{X} \sim N\left(610, \frac{22^2}{4}\right) = N(610, 11^2).$$

Standardise:

Lower:

$$Z_1 = \frac{600 - 610}{11} = \frac{-10}{11} \approx -0.909.$$

Upper:

$$Z_2 = \frac{620 - 610}{11} = \frac{10}{11} \approx 0.909.$$

So

$$P(600 < \bar{X} < 620) = P(-0.909 < Z < 0.909).$$

From tables, $P(Z < 0.909) \approx 0.818$, so

$$P(-0.909 < Z < 0.909) \approx 0.818 - (1 - 0.818) = 0.818 - 0.182 = 0.636.$$

Question 3 — Radon Levels

Radon levels: $X \sim N(\mu, 1.5^2)$, sample of $n = 6$:

$$18.5, 19.1, 21.7, 20.3, 18.4, 21.4.$$

Compute sample mean:

$$\sum = 18.5 + 19.1 + 21.7 + 20.3 + 18.4 + 21.4 = 119.4.$$

$$\bar{x} = \frac{119.4}{6} = 19.9.$$

Standard error:

$$\text{SE} = \frac{\sigma}{\sqrt{n}} = \frac{1.5}{\sqrt{6}} \approx \frac{1.5}{2.45} \approx 0.612.$$

For a 98% confidence interval, $z \approx 2.33$.
Margin of error:

$$E = z \times \text{SE} \approx 2.33 \times 0.612 \approx 1.43.$$

So the 98% confidence interval is:

$$\mu \in (19.9 - 1.43,\ 19.9 + 1.43) \approx (18.5,\ 21.3).$$

Question 4 — Orange Weights

Known $\sigma = 10$ g, sample size $n = 12$, sample mean $\bar{x} = 193$ g. We want a 99% confidence interval for μ.

Standard error:

$$\text{SE} = \frac{10}{\sqrt{12}} \approx \frac{10}{3.464} \approx 2.89.$$

For 99%, $z \approx 2.576$.
Margin of error:

$$E = 2.576 \times 2.89 \approx 7.44.$$

Confidence interval:

$$\mu \in (193 - 7.44,\ 193 + 7.44) \approx (185.6,\ 200.4).$$

Comment: The supplier's claim $\mu = 200$ g lies just inside this interval, so it is plausible based on this sample.

Question 5 — IQ Scores

IQ $\sim N(\mu, 20^2)$.

Couple averages:

$$94,\ 118,\ 129,\ 103,\ 134,\ 92,\ 99,\ 95.$$

(a) Show $N(108, 20^2)$ Compute mean of the eight averages:

$$\sum = 94 + 118 + 129 + 103 + 134 + 92 + 99 + 95 = 864.$$

$$\bar{x} = \frac{864}{8} = 108.$$

The student has taken this as an estimate of μ, and kept the given standard deviation 20. So the estimated distribution is $N(108, 20^2)$.

(b) Assumptions

- The eight couples are a random, representative sample of the town's adult population.
- The average of each couple is an unbiased estimate of the population mean.
- The population standard deviation is known and equal to 20.

(c) Comment on assumptions

- Using only eight couples is a small sample; sampling variability is high.
- Couples may not be representative (e.g. similar education, socio-economic status).
- Treating the given standard deviation as fixed may be unrealistic if it was only estimated.

Overall, the estimate $N(108, 20^2)$ is rough and based on strong assumptions; it should be treated with caution.

5. Tesla Daily Returns

The daily returns on Tesla Inc. (TSLA) stock have a mean of 0.05% and a standard deviation of 2.1%.

> **Note:** These figures are based upon historic performance over the last 10 years and are not guaranteed for the future.

(a) Comment on the significance of the value of the mean.

(b) Comment on the value of the standard deviation compared to that of the mean.

(c) Assume that the daily returns are normally distributed.

  (i) On how many days in a year would you expect the daily return to be less than −2.05%?

  (ii) On how many days in a year would you expect the daily return to be greater than 4.25%?

Mark Scheme — Tesla Daily Returns

(a) Significance of the mean The mean daily return of 0.05% indicates that, on average, Tesla's stock price increases slightly each day. This suggests long-term growth, but the value is small compared to daily volatility.

(b) Comparison of standard deviation to mean The standard deviation is 2.1%, which is much larger than the mean of 0.05%. This implies that daily returns fluctuate significantly and that the average return is not a reliable predictor of short-term performance.

(c)(i) Probability of return less than -2.05% Let $X \sim N(0.05, 2.1^2)$.

Standardise:

$$Z = \frac{-2.05 - 0.05}{2.1} = \frac{-2.10}{2.1} = -1.00$$

From standard normal tables:

$$P(Z < -1.00) \approx 0.159$$

Assuming 252 trading days in a year:

$$252 \times 0.159 \approx 40.1$$

Answer: Approximately 40 days per year.

(c)(ii) Probability of return greater than 4.25% Standardise:

$$Z = \frac{4.25 - 0.05}{2.1} = \frac{4.20}{2.1} = 2.00$$

From standard normal tables:

$$P(Z > 2.00) = 1 - P(Z < 2.00) \approx 1 - 0.977 = 0.023$$

Expected days:

$$252 \times 0.023 \approx 5.8$$

Answer: Approximately 6 days per year.

Correlation and Regression — Shadow Questions

1. Ice Cream Van Sales

An ice cream van owner records the daily temperature (x, °C) and the day's takings (y, £) for 8 randomly chosen summer days.

Temperature x	16	18	21	23	25	27	29	31
Takings y (£)	62	78	95	108	121	140	158	176

(a) Describe the correlation shown by this data.

(b) The PMCC for this data is $r = 0.995$. State what this value tells you, and comment on whether it proves that hot weather causes higher takings.

(c) A local newspaper claims that global ice cream sales rising alongside global temperatures over the last 50 years proves climate change is being caused by the ice cream industry. Explain the flaw in this reasoning.

2. House Prices

A regression line relating the floor area of a house, x (m^2, ranging from 40 to 160 in the sample), to its sale price, y (£1000s), is

$$y = 45 + 2.6x$$

(a) Predict the sale price of a house with floor area $90\,\text{m}^2$.

(b) Predict the sale price of a house with floor area $500\,\text{m}^2$, and explain why this prediction should not be trusted.

(c) Interpret the value 2.6 in context.

3. Revision and Exam Performance

A teacher records the number of hours of revision, x, and the exam score, y (%), for 10 students, with x ranging from 3 to 22 hours. The PMCC is $r = 0.87$ and the regression line of y on x is $y = 41 + 2.1x$.

(a) Interpret the value of r.

(b) Predict the score of a student who revised for 15 hours.

(c) One student revised for 22 hours and scored only 58%. Suggest one reason why this student's actual score might not match the value predicted by the regression line.

4. Fitness Training

A sports scientist studies the relationship between the number of weeks of training, x (ranging from 0 to 16), and the time, y seconds, taken by an athlete to run 1500 m. The regression line of y on x is

$$y = 320 - 3.5x$$

(a) Predict the athlete's 1500 m time after 10 weeks of training.

(b) Explain why using this line to predict the athlete's time after 5 years of training would not be sensible.

(c) Interpret the gradient -3.5 in context.

5. Broadband Speed and Streaming Complaints

A telecoms regulator collects data from 40 regions on average broadband speed (x, Mbps) and the number of streaming-quality complaints received per 1000 households (y). The PMCC is found to be $r = -0.79$.

(a) Describe the strength and direction of this correlation.

(b) A regional manager claims: "This proves that upgrading broadband speed will directly reduce complaints in every region." Comment on this claim, suggesting a possible confounding variable.

(c) Explain why a value of r close to -1 or $+1$ does not, by itself, guarantee that a regression line will give reliable predictions.

Mark Scheme — Correlation and Regression

Question 1 — Ice Cream Van Sales

(a) As temperature increases, takings increase – this is strong positive correlation.

(b) $r = 0.995$ is very close to $+1$, indicating very strong positive linear correlation between temperature and takings. It does *not* prove causation on its own – although here a direct causal mechanism (hot weather makes people want ice cream) is highly plausible, the correlation alone is not sufficient proof.

(c) This is a correlation-does-not-imply-causation error at a global scale: rising ice cream sales and rising global temperatures could both be driven by other factors (population growth, economic growth increasing disposable income) rather than one directly causing the other. The direction of causation is also reversed from what would make physical sense – ice cream sales cannot cause climate change through this mechanism.

Question 2 — House Prices

(a) $y = 45 + 2.6 \times 90 = 279$, so the predicted price is £279,000.

(b) $y = 45 + 2.6 \times 500 = 1345$, i.e. £1,345,000. This is extrapolation, since $500\,\mathrm{m}^2$ is far outside the data range (40–$160\,\mathrm{m}^2$), so the linear relationship cannot be assumed to continue and the prediction should not be trusted.

(c) For every extra square metre of floor area, the sale price is estimated to increase by £2,600, according to this model.

Question 3 — Revision and Exam Performance

(a) $r = 0.87$ is close to $+1$, indicating strong positive linear correlation – students who revise more tend to score higher.

(b) $y = 41 + 2.1 \times 15 = 72.5\%$.

(c) Any sensible factor affecting an individual's performance beyond revision time, e.g. exam anxiety, quality (not just quantity) of revision, an off day, or a lack of sleep – the regression line predicts an *average* trend, not every individual data point exactly.

Question 4 — Fitness Training

(a) $y = 320 - 3.5 \times 10 = 285$ seconds.

(b) 5 years (260 weeks) is far outside the data range (0–16 weeks) – this is extreme extrapolation. The model would predict an impossibly fast (or even negative) time, since performance cannot keep improving linearly forever; there is a physical limit to how fast a human can run.

(c) For every additional week of training, the athlete's predicted 1500 m time decreases by 3.5 seconds.

Question 5 — Broadband Speed and Streaming Complaints

(a) $r = -0.79$ indicates a fairly strong negative linear correlation: regions with higher broadband speeds tend to have fewer streaming complaints.

(b) The claim overstates what correlation shows. A plausible confounder is regional wealth/urbanisation: wealthier or more urban regions may have both better broadband infrastructure *and* fewer complaints for unrelated reasons (e.g. newer streaming equipment, greater consumer awareness of how to troubleshoot issues), rather than speed alone driving the complaint rate down in every region.

(c) A strong PMCC only confirms a strong *linear* relationship exists in the data used to calculate it; it does not guarantee that the regression line will predict accurately for individual cases, for regions outside the range of the data (extrapolation), or that the relationship will remain linear or stable over time.