

BCB744 Biostatistics – Theory Test (Version 9)

Total: 135 marks | Time: 180 minutes

A. J. Smit
University of the Western Cape

2026-01-01

! Important

Instructions

- This paper has **three parts**: Part A (General Theory, 61 marks), Part B (Experiment Design and Hypothesis Formulation, 37 marks), and Part C (Statistical Output Interpretation, 37 marks).
- Mark allocations are shown next to each question in (/ **marks**) notation.
- Answer **all** questions.
- Write clearly and in complete sentences where prose is required.
- Number all questions clearly and use the Quarto headings facility to assign the main question number to level 1 (e.g., # **Question 1**) and the subordinate parts to level 2 (e.g., ## **Q1a**).
- Statistical notation: use H_0 for the null hypothesis and H_A for the alternative hypothesis.
- You are **not** allowed access to the internet or AI.
- You **may** use the cheatsheet and the RStudio/R help files.
- You **must** submit your knitted document in `.html` format on iKamva immediately after the 3-hr test duration has elapsed.
- Use `embed-resources: true` in Quarto's YAML header to ensure the `.html` file displays correctly.
- **Any** format other than `.html` will be disqualified from assessment.

Part A: General Theory (61 marks)

Question 1 – Variables and Measurement Scales (/6)

- a. Describe the **four levels of measurement** (nominal, ordinal, interval, ratio). For each level, give one biological example. (/ 4)
- b. Why does the level of measurement of a response variable **constrain the choice of statistical test**? Give one concrete example where using a test designed for a higher measurement level on a lower-level variable would be problematic. (/ 2)

 Tip

Model Answer – Question 1

a. One mark per level with a valid example:

- ✓ **Nominal**: categories with no inherent order; differences have no quantitative meaning. Example: species identity (damselfish, parrotfish, wrasse) or habitat type (rocky shore, sandy beach, seagrass bed).
- ✓ **Ordinal**: categories with a meaningful rank order, but intervals between ranks are not equal. Example: substrate rugosity scored 1–5 (low to high), or dominance rank in a social group.
- ✓ **Interval**: continuous measurements with equal intervals between values, but an arbitrary zero (zero does not mean absence). Example: water temperature in °C (0°C is arbitrary – does not mean absence of heat).
- ✓ **Ratio**: continuous measurements with a true zero (zero means complete absence of the quantity). Example: body mass (g), shell length (mm), or dissolved oxygen (mg L^{-1}) – a value of zero means none is present.

b.

- ✓ Parametric tests (e.g., ANOVA, t -tests) require at least interval-level measurement, because they use arithmetic operations (mean, variance) that assume equal spacing between values. Applying these tests to ordinal data assumes equal intervals that do not exist.
- ✓ Example: calculating the mean of an ordinal rugosity score (1–5) treats the difference between 1 and 2 as equal to the difference between 4 and 5 – which is not guaranteed. A non-parametric test (e.g., Kruskal-Wallis) is appropriate for ordinal response variables, as it operates on ranks rather than raw values.

Question 2 – Sampling Distributions and the Central Limit Theorem (/7)

- a. State the **Central Limit Theorem (CLT)**. Why is it important for applying parametric hypothesis tests to biological data that are not perfectly normally distributed? (/ 3)
- b. A researcher takes repeated samples of $n = 30$ fish from a large lake and records the mean body length each time. Describe the **shape**, **centre**, and **spread** of the resulting distribution of sample means. (/ 2)
- c. Show mathematically why increasing the sample size from $n = 25$ to $n = 100$ halves the width of the 95% confidence interval (assume SD remains constant). (/ 2)

 Tip

Model Answer – Question 2

a.

- ✓ The **Central Limit Theorem** states that the sampling distribution of the mean of a sufficiently large random sample will be approximately normally distributed, regardless of the shape of the underlying population distribution. The approximation improves as sample size (n) increases.
- ✓ This is important because it means that even when individual measurements are skewed, bimodal, or otherwise non-normal, the means of large samples will follow a normal distribution – validating parametric tests (which require normally distributed test statistics or estimators) without requiring perfect normality of raw data.
- ✓ A common rule of thumb is that $n \geq 30$ is sufficient for the CLT to apply in most biological contexts, though symmetric distributions converge faster than heavily skewed ones.

b.

- ✓ **Shape:** approximately normal (by the CLT, since $n = 30$ is adequate for most biological variables).
- ✓ **Centre:** the mean of the sampling distribution equals the true population mean (μ) – the sample mean is an unbiased estimator.
- ✓ **Spread:** the standard deviation of the sampling distribution (the standard error) = $SD / \sqrt{n} = SD / \sqrt{30}$. It is much smaller than the population SD because each sample mean averages out individual variation.

c.

- ✓ The 95% CI half-width $\approx t_{\text{crit}} \times SE = t_{\text{crit}} \times (SD / \sqrt{n})$. Since SD and t_{crit} are constant, CI width $\propto 1 / \sqrt{n}$.
- ✓ At $n = 25$: width $\propto 1 / \sqrt{25} = 1/5$. At $n = 100$: width $\propto 1 / \sqrt{100} = 1/10$. Ratio = $(1/10) / (1/5) = 1/2$ – the CI width at $n = 100$ is exactly half that at $n = 25$. Quadrupling the sample size halves the CI width.

Question 3 – Poisson and Negative Binomial Distributions (/7)

- a. State **three conditions** under which count data are expected to follow a Poisson distribution. What is the key equidispersion property, and how would you check it? (/ 3)
- b. When counts show **overdispersion** (variance > mean), the negative binomial distribution is preferred over Poisson. What does overdispersion indicate biologically, and why is it so common in ecological count data? (/ 2)
- c. A researcher counts spider crabs on 80 rocky reef patches. The sample mean is 3.2 crabs per patch and the sample variance is 8.7. Is a Poisson distribution appropriate? Quantify the degree of overdispersion. (/ 2)

💡 Tip

Model Answer – Question 3

a.

- ✓ (i) Events (crab sightings) occur **independently** – the presence of one crab does not affect the probability of another being present. (ii) The **rate** of occurrence is **constant** across the sampling area or time interval – all patches have the same underlying probability of a crab being present. (iii) Only one event can occur at any given instant (no simultaneous multiple events), and the count per patch is a **non-negative integer**.
- ✓ The key equidispersion property: for a Poisson distribution, the **variance equals the mean** ($\text{Var}(X) = \lambda = E(X)$). Check it by comparing the sample variance to the sample mean; a dispersion ratio (variance/mean) close to 1 supports Poisson; ratios well above 1 indicate overdispersion.

b.

- ✓ Overdispersion indicates that organisms or events are **more clustered** (aggregated) than would be expected under random (independent) occurrence – individual patches contain far more or far fewer crabs than expected if they were randomly distributed. This is the rule rather than the exception in ecology: animals aggregate around food sources, shelter, or conspecifics; rare events cluster in space or time.
- ✓ Ecologically, overdispersion commonly arises from **habitat heterogeneity** (some patches are better quality), **conspecific attraction** (crabs aggregate together), or **environmental patchiness** (current or food concentration). The negative binomial explicitly models this extra variation via a dispersion parameter.

c.

- ✓ The **dispersion ratio** = variance / mean = $8.7 / 3.2 \approx 2.72$ – far greater than 1. A Poisson distribution is **not appropriate** for these data.
- ✓ The degree of overdispersion is 2.72-fold: the observed variance is almost three times the variance expected under Poisson. This indicates strong aggregation of spider crabs across patches, and a **negative binomial distribution** would be a more appropriate model.

Question 4 – One-Sample *t*-Test (/6)

- a. When is a **one-sample *t*-test** appropriate? What quantity does it compare, and against what reference? (/ 2)
- b. A physiologist measures blood pH of 15 individual rock lobsters (*Jasus lalandii*) and tests whether the mean pH differs from the physiological reference value of 7.82. State the appropriate H_0 and H_A , and identify the quantity that enters the test statistic. (/ 2)
- c. The test returns $t(14) = -2.54$, $p = 0.024$. Interpret this result fully at $\alpha = 0.05$. (/ 2)

 Tip

Model Answer – Question 4

a.

- ✓ A one-sample t -test is appropriate when you have a **single sample of continuous measurements** and wish to test whether the population mean differs from a specific **reference or theoretical value** (μ_0) that is known from prior knowledge, physiology, or regulation (not estimated from the data).
- ✓ It compares the **sample mean** (\bar{x}) against the hypothesised population mean (μ_0) via the test statistic $t = (\bar{x} - \mu_0) / (SE) = (\bar{x} - \mu_0) / (SD / \sqrt{n})$.

b.

- ✓ H_0 : The mean blood pH of *J. lalandii* equals the physiological reference value; $\mu = 7.82$.
- ✓ H_A : The mean blood pH differs from the reference; $\mu \neq 7.82$ (two-tailed, since we have no a priori reason to predict the direction of deviation).
- ✓ The quantity that enters the test statistic is the **difference between the sample mean and 7.82**, scaled by the standard error: $t = (\bar{x} - 7.82) / (SD / \sqrt{15})$.

c.

- ✓ $t(14) = -2.54$, $p = 0.024 < \alpha = 0.05$: we **reject** H_0 . There is statistically significant evidence that the mean blood pH of rock lobsters differs from the reference value of 7.82.
- ✓ The negative t -value indicates that the sample mean is **below** 7.82 — the lobsters' blood is more acidic than the reference value. The probability of observing a mean this far from 7.82 (in either direction) by chance, if the true mean were 7.82, is only 2.4%.

Question 5 – Type I and Type II Errors: Balancing Risk (/6)

- a. Define **Type I** and **Type II errors**. Which one does the significance level α directly control, and how? (/ 2)
- b. A conservation biologist sets $\alpha = 0.10$ (rather than the conventional 0.05) to test whether a reintroduced cheetah population has established. What is their reasoning, and what statistical risk do they accept by doing this? (/ 2)
- c. If 50 independent tests are each conducted at $\alpha = 0.05$ when all null hypotheses are true, how many false positives would you **expect**, and why? (/ 2)

💡 Tip

Model Answer – Question 5

a.

- ✓ A **Type I error** (false positive) is rejecting a true H_0 – concluding there is an effect when none exists. A **Type II error** (false negative) is failing to reject a false H_0 – missing a real effect.
- ✓ α directly controls the **Type I error rate**: setting $\alpha = 0.05$ means we accept a 5% probability of falsely rejecting a true H_0 . It sets the critical threshold for the p -value below which we reject H_0 .

b.

- ✓ **Reasoning**: in a conservation context, a Type II error (failing to detect an established population and wrongly concluding the reintroduction failed) may be more costly than a Type I error – ending a successful conservation programme prematurely has severe ecological consequences. By raising α to 0.10, the biologist makes the test more *sensitive* (easier to detect a real effect), reducing the risk of missing a true establishment.
- ✓ **Risk accepted**: by increasing α , they increase the Type I error rate to 10% – a 10% chance of declaring the population established when it is not, potentially wasting resources on an ineffective reintroduction programme.

c.

- ✓ Expected false positives = number of tests $\times \alpha = 50 \times 0.05 = 2.5$ **false positives** expected.
- ✓ Each test has a 5% chance of a Type I error when H_0 is true. With 50 independent tests, the expected number of spurious rejections is 2.5 – on average, approximately 2–3 of the 50 “significant” results are false positives arising purely by chance, with no real effect. This is the essence of the multiple comparisons problem.

Question 6 – Pearson vs Spearman Correlation (/7)

- a. Explain the **conceptual difference** between Pearson’s r and Spearman’s ρ (rho). Under what circumstances would you choose Spearman over Pearson? (/ 3)
- b. A survey of 30 rocky-shore transects finds $r = 0.88$ between barnacle cover (%) and mean daily wave height (m). A journal reviewer argues that this result demonstrates that “wave height directly controls barnacle cover.” Give **two specific reasons** why this interpretation is problematic, and explain what additional evidence would be needed to support a causal claim. (/ 4)

💡 Tip

Model Answer – Question 6

a.

- ✓ **Pearson's r** measures the strength and direction of the **linear** association between two continuous variables. It is sensitive to outliers and requires that both variables are approximately normally distributed and that the relationship is linear.
- ✓ **Spearman's ρ** measures the **monotonic** association between the **ranks** of two variables – it assesses whether values tend to increase together, but without assuming a linear or normally distributed relationship. It is robust to outliers and applicable to ordinal data.
- ✓ Choose Spearman when: the relationship is monotonic but non-linear; the data contain outliers that would distort Pearson's r ; one or both variables are ordinal; or the assumption of bivariate normality is clearly violated.

b.

- ✓ **Confounding**: both wave height and barnacle cover may be influenced by a third variable – for example, site exposure (south-facing vs. north-facing) may independently drive both high wave energy and high barnacle settlement, producing a correlation that does not reflect a direct mechanistic link.
- ✓ **Direction of causality**: correlation is symmetric – it quantifies co-variation but cannot determine whether wave height drives barnacle cover, barnacle cover (by altering surface roughness) affects local wave energy, or both are driven by a common upstream factor. Observational data cannot establish the direction of causation.
- ✓ To support causation: one would need a **manipulative experiment** (e.g., wave-exposure cages that alter local wave energy on reefs while keeping other factors constant), or at minimum a mechanistic pathway (e.g., wave disturbance removes settled barnacles, and the mechanism has been demonstrated in laboratory conditions and field manipulations). A strong correlation alone, however high, is insufficient.

Question 7 – Residual Diagnostic Plots (/7)

For each of the following patterns observed in regression diagnostic plots, describe (i) what assumption is violated, (ii) the biological or statistical implication, and (iii) one corrective action.

- a. A **fan-shaped spread** in the residuals-vs-fitted plot, where residuals widen as fitted values increase. (/ 2)
- b. A **U-shaped curve** in the residuals-vs-fitted plot. (/ 2)
- c. Points that **curve away from the reference line** at both ends of a Q-Q plot, bending upward at the top-right and downward at the bottom-left. (/ 2)
- d. After applying a log-transformation to the response, the residuals-vs-fitted plot shows random scatter and the Q-Q plot is approximately straight. What does this tell you? (/ 1)

💡 Tip

Model Answer – Question 7

a.

- ✓ Assumption violated: **homoscedasticity** (constant variance). Residual variance increases with the fitted value – the model errors are larger for higher predicted values.
- ✓ Implication: standard errors are underestimated at high fitted values and overestimated at low values, making t - and F -tests unreliable.
- ✓ Corrective action: apply a **variance-stabilising transformation** to the response (log or square-root); or use weighted least squares.

b.

- ✓ Assumption violated: **linearity**. The systematic U-shape means the linear model under-predicts at low and high fitted values, and over-predicts in the middle – the true relationship is curved.
- ✓ Implication: the linear model is mis-specified; predictions across the range will be systematically biased.
- ✓ Corrective action: add a **quadratic (polynomial) term** to the predictor; or transform the predictor variable.

c.

- ✓ Assumption violated: **normality of residuals** – the S-curve (bending away from the line at both ends) indicates **heavy tails** (leptokurtosis): more extreme residuals than expected under a normal distribution.
- ✓ Implication: p -values derived from normal-theory tests may be unreliable; extreme observations have more influence than the model assumes.
- ✓ Corrective action: investigate extreme residuals for data errors; consider a log or Box-Cox transformation; use a non-parametric test.

d.

- ✓ The log-transformation has **successfully resolved** both non-constant variance and non-normality in the residuals. The assumptions of the linear model are now sufficiently met, and the parametric analysis can proceed on the log-transformed scale. The relationship is approximately **log-linear**: a linear model on the log scale is equivalent to an exponential model on the original scale.

Question 8 – Simple Linear Regression: Full Interpretation (/8)

- a. A regression of seagrass shoot density (shoots m^{-2}) on water depth (m) returns: $\hat{y} = 145.2 - 12.4x$, $R^2 = 0.63$. Interpret the **slope**, **intercept**, and R^2 in biological terms. (/ 4)
- b. Distinguish between a **confidence interval** and a **prediction interval** for a regression model. Which is wider, and why? (/ 2)

- c. The regression is re-run with depth values **centred** at the grand mean depth of 4.5 m. The new intercept is 89.0. Verify that this is consistent with the original equation and explain what the centred intercept represents. (/ 2)

 Tip

Model Answer – Question 8

a.

- ✓ **Slope (-12.4)**: for each 1 m increase in water depth, shoot density is predicted to decrease by **12.4 shoots m⁻²**. The negative slope is biologically consistent with light attenuation – deeper water reduces the photosynthetically active radiation reaching the canopy, limiting seagrass growth.
- ✓ **Intercept (145.2)**: the predicted shoot density when depth = 0 m (at the surface). While mathematically defined, this value represents an extrapolation to the waterline edge and may not be biologically meaningful unless sampling included very shallow sites.
- ✓ $R^2 = 0.63$ means that **63% of the variation in shoot density** among sampling locations is explained by variation in water depth. The remaining 37% is attributable to other factors (e.g., substrate type, nutrient availability, herbivory) not captured by the model.

b.

- ✓ A **confidence interval** (CI) for a fitted value represents the uncertainty in the **mean** response at a given x – the range of plausible values for the true population mean of y at that depth.
- ✓ A **prediction interval** (PI) represents the uncertainty for a **single new observation** at a given x – it incorporates both the uncertainty in the mean line (the CI component) and the natural scatter of individual observations around that mean (residual variance).
- ✓ The PI is always wider because it must account for two sources of uncertainty: (1) the position of the mean line, and (2) the inherent variability of individual shoots around the line.

c.

- ✓ Substituting $x = 4.5$ m into the original equation: $\hat{y} = 145.2 - 12.4 \times 4.5 = 145.2 - 55.8 = \mathbf{89.4} \approx 89.0$ ✓ (small rounding difference).
- ✓ The centred intercept (89.0) represents the **predicted shoot density at the mean depth** (4.5 m) – a biologically meaningful and directly observable quantity, unlike the original intercept at depth = 0. Centring does not change the slope or fit; it simply relocates the intercept to the centre of the data cloud, making it interpretable.

Question 9 – Polynomial and Multiple Regression: Model Comparison
(/7)

- a. Explain why polynomial regression is described as “**linear in its parameters**” even though it models a curved relationship. (/ 2)
- b. A researcher fits both a linear and a quadratic model to data on barnacle density vs. intertidal height. The models return:

Model	AIC	Adjusted R^2
Linear	312.4	0.48
Quadratic	298.7	0.67

Which model is preferred? Justify using **both** criteria. (/ 2)

- c. What is **overfitting** in the context of polynomial regression? How does the adjusted R^2 protect against it, compared to R^2 ? (/ 3)



Tip

Model Answer – Question 9

a.

- ✓ A model is **linear in its parameters** if each coefficient (β) appears as a simple multiplier – it is never raised to a power, placed inside a non-linear function, or multiplied by another coefficient. In $\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2$, the parameters β_0 , β_1 , and β_2 each enter linearly as multipliers of 1, x , and x^2 respectively.
- ✓ The curvature arises from the predictor term x^2 – which is simply a new column in the design matrix. The estimation (via ordinary least squares) is identical to multiple linear regression; only the predictor, not the model structure, is non-linear.

b.

- ✓ The **quadratic model is preferred** on both criteria: it has a **lower AIC** (298.7 vs. 312.4; $\Delta\text{AIC} = 13.7$, well above the threshold of 10) indicating substantially better fit per unit of complexity; and a **higher adjusted R^2** (0.67 vs. 0.48), indicating the quadratic term improves fit more than expected by chance after penalising for the additional parameter.
- ✓ The AIC difference of 13.7 provides **strong support** for the quadratic model – the linear model is not competitive. The adjusted R^2 increase of 0.19 is also practically meaningful, confirming the quadratic term captures real curvature in the barnacle density–height relationship.

c.

- ✓ **Overfitting** occurs when a polynomial model is too complex relative to the data – it fits the noise of the sample (idiosyncratic fluctuations) rather than the true underlying relationship, so it performs well on the training data but poorly on new data.
- ✓ R^2 always increases (or stays the same) whenever a term is added, regardless of whether it captures real signal or just noise – it cannot detect overfitting.
- ✓ **Adjusted R^2** penalises each additional parameter by the factor $(n - 1) / (n - k - 1)$. If a term adds less explanatory power than expected by chance, adjusted R^2 **decreases** – it provides a built-in protection against overfitting by declining when unnecessary complexity is added.

Part B: Experiment Design and Hypothesis Formulation (37 marks)

Question 10 – Two-Way ANOVA: Seagrass Photosynthesis (/12)

A marine botanist tests the effect of **light intensity** (three levels: Low, Medium, High) and **CO₂ concentration** (two levels: Ambient, Elevated) on the net photosynthesis rate ($\mu\text{mol CO}_2 \text{ m}^{-2} \text{ s}^{-1}$)

of seagrass (*Zostera capricorni*). Four individual plants are assigned to each of the six treatment combinations. The first six rows of the dataset are:

plant_id	light	co2	net_photo
1	1	Low Ambient	2.14
2	2	Low Ambient	1.89
3	3	Low Elevated	2.73
4	4	Low Elevated	2.51
5	5	Medium Ambient	5.42
6	6	Medium Ambient	5.17

The researcher asks: “Does light intensity, CO₂ concentration, or their interaction affect seagrass net photosynthesis?”

- State formal null and alternative hypotheses for (i) the main effect of CO₂ concentration and (ii) the light × CO₂ interaction. (/ 4)
- Identify the most appropriate statistical test and give **three reasons** for your choice. (/ 4)
- The interaction term is **significant**. Explain what this means biologically and what consequence it has for how you interpret and report the main effect of light intensity. (/ 2)
- If the main effect of light intensity is significant, what post-hoc procedure would you apply, and how many pairwise comparisons does it involve? (/ 2)

💡 Tip

Model Answer – Question 10

a. Two marks per effect ($H_0 + H_A$):

(i) CO_2 main effect:

- ✓ H_0 : Mean net photosynthesis does not differ between ambient and elevated CO_2 , averaging across all light levels ($\mu_{\text{Ambient}} = \mu_{\text{Elevated}}$).
- ✓ H_A : Mean net photosynthesis differs between CO_2 concentrations ($\mu_{\text{Ambient}} \neq \mu_{\text{Elevated}}$).

(ii) Light \times CO_2 interaction:

- ✓ H_0 : The effect of CO_2 concentration on net photosynthesis is the same at all three light levels – the factors act additively and independently.
- ✓ H_A : The effect of CO_2 concentration on net photosynthesis depends on light intensity – the two factors interact non-additively.

b.

- ✓ **Two-way (factorial) ANOVA.**
- ✓ Reason 1: There are **two categorical predictors** (light: 3 levels; CO_2 : 2 levels), each with distinct levels. Two-way ANOVA simultaneously tests both main effects and their interaction – a design not accommodated by one-way ANOVA or t -tests.
- ✓ Reason 2: The response (net photosynthesis rate, $\mu\text{mol } CO_2 \text{ m}^{-2} \text{ s}^{-1}$) is **continuous** and ratio-scaled, appropriate for ANOVA which compares group means.
- ✓ Reason 3: The design is **balanced** (4 plants per cell), which maximises the power and computational simplicity of the factorial ANOVA – all cells are estimated with equal precision.

c.

- ✓ A significant interaction means that the **effect of CO_2 on photosynthesis depends on the light level** (or equivalently, the response to increasing light differs between ambient and elevated CO_2). For example, elevated CO_2 may boost photosynthesis much more at high light (where CO_2 is the limiting factor) than at low light (where light itself is limiting regardless of CO_2).
- ✓ Because the interaction is significant, the **main effect of light intensity cannot be interpreted in isolation** – “light increases photosynthesis by X” is misleading if this effect differs markedly between CO_2 levels. The interaction must be interpreted first, and the light effect reported separately for each CO_2 level (simple main effects), ideally using an interaction plot.

d.

- ✓ Apply **Tukey’s Honestly Significant Difference (HSD)** test to identify which specific light levels differ.
- ✓ With 3 light levels (Low, Medium, High), the number of pairwise comparisons = $3(3-1)/2 = 3$ (Low vs. Medium, Low vs. High, Medium vs. High).

Question 11 – Multiple Regression: Octopus Arm Span (/13)

A marine biologist measures the arm span (cm) of 62 individual common octopuses (*Octopus vulgaris*) along with three continuous predictors: mantle length (mm), age (months), and prey availability index (PAI; dimensionless, 0–10 scale). The first six rows of the dataset are:

	octopus_id	arm_span_cm	mantle_mm	age_months	prey_index
1	1	74.3	142.1	8.4	6.2
2	2	81.7	158.4	9.8	5.7
3	3	68.2	131.8	7.1	4.9
4	4	92.4	174.6	11.2	7.3
5	5	65.8	127.3	6.8	3.4
6	6	88.1	169.2	10.5	6.8

The research question is: “Which environmental and morphological variables best predict arm span in *O. vulgaris*?”

- State the null and alternative hypotheses for the **overall** multiple regression model. (/ 3)
- What does a **partial regression coefficient** represent in this model? How does it differ from the slope in a simple regression of arm span on mantle length alone? (/ 3)
- Inspecting the data, mantle length and age appear strongly correlated. How would you diagnose **collinearity** between these predictors, and what would you do if it is confirmed? (/ 4)
- How does **adjusted R^2** help evaluate whether all three predictors should be retained in the model? (/ 3)



Tip

Model Answer – Question 11

a.

- ✓ H_0 : None of the three predictors (mantle length, age, prey index) has a linear relationship with arm span; all partial slopes $\beta_1 = \beta_2 = \beta_3 = 0$. The model explains no more variance than the intercept-only null model.
- ✓ H_A : At least one predictor has a non-zero partial slope – at least one variable is a significant linear predictor of arm span.
- ✓ The overall null is tested by the omnibus F -statistic in the ANOVA table of the regression output.

b.

- ✓ A **partial regression coefficient** for mantle length estimates the change in arm span associated with a one-unit increase in mantle length, **holding age and prey index constant**. It represents the unique contribution of mantle length to arm span, statistically isolated from the variation shared with the other two predictors.
- ✓ In a **simple regression** of arm span on mantle length only, the slope captures the total association – it conflates the direct effect of mantle length with any variation shared between mantle length and age (since larger octopuses are also older, and age independently predicts arm span). The simple slope is biased by the unmeasured confounders that are correlated with mantle length.

c.

- ✓ Diagnose collinearity using the **Variance Inflation Factor (VIF)**: $VIF_j = 1 / (1 - R_j^2)$, where R_j^2 is the variance in predictor j explained by all other predictors. A $VIF > 5$ (or 10 by more lenient standards) for mantle length or age indicates problematic collinearity.
- ✓ If confirmed: (i) **remove one of the collinear predictors** – typically the one with less theoretical justification (e.g., remove age if mantle length captures the same information more directly); or (ii) combine the correlated predictors into a single composite variable (e.g., a principal component); or (iii) centre the predictors (does not resolve collinearity but improves interpretability of the intercept).

d.

- ✓ Adjusted R^2 penalises each predictor added: it increases only when the new predictor adds more explanatory power than expected by chance. If removing one predictor (e.g., prey index) from the three-predictor model **raises** adjusted R^2 , it suggests that predictor was adding noise, not signal, and should be excluded.
- ✓ Conversely, if adjusted R^2 drops substantially when a predictor is removed, that predictor is contributing genuine explanatory power. The preferred model is the one with the highest adjusted R^2 (equivalently, the lowest AIC) among plausible candidate models – not necessarily the model with the most predictors.

Question 12 – Polynomial Regression: Barnacle Density Along a Tidal Gradient (/12)

An ecologist samples 45 quadrats (25 cm²) at different intertidal heights (m above chart datum, ranging from 0.2 to 2.8 m) and counts barnacle density (individuals per quadrat). The data preview is:

quadrat	height_m	barnacle_density	
1	1	0.2	8
2	2	0.4	14
3	3	0.6	22
4	4	0.8	31
5	5	1.0	44
6	6	1.2	51
...			
20	20	1.8	58
...			
40	40	2.4	28
45	45	2.8	12

The researcher suspects the relationship is **unimodal** (barnacle density peaks at mid-tidal height and declines both lower and higher).

- Formulate appropriate null and alternative hypotheses for testing whether a **quadratic model** provides a significantly better fit than a linear model. (/ 3)
- Explain why a linear model is biologically inappropriate here, with reference to the data preview. (/ 3)
- Describe **two methods** to determine whether the quadratic term significantly improves model fit. (/ 3)
- What biological risk does **overfitting** with a high-degree polynomial pose for ecological interpretation of this gradient? (/ 3)

💡 Tip

Model Answer – Question 12

a.

- ✓ H_0 : The quadratic term (height²) does not significantly improve the fit over the linear model; the coefficient $\beta_2 = 0$. A linear model adequately describes the barnacle–height relationship.
- ✓ H_A : The quadratic term significantly improves model fit; $\beta_2 \neq 0$. The true relationship is curvilinear (hump-shaped).
- ✓ The test compares the nested models (linear \subset quadratic) and can be stated formally as: $H_0: \beta_2 = 0$ in the model $\hat{y} = \beta_0 + \beta_1x + \beta_2x^2$.

b.

- ✓ The data preview shows barnacle density **increasing** from 8 at 0.2 m to 58 at 1.8 m, then **declining** back to 12 at 2.8 m. A linear model cannot represent a response that first rises then falls – it would fit a single-direction slope and systematically under-predict at intermediate heights and over-predict at both extremes.
- ✓ Biologically, this hump-shaped pattern is expected: barnacles at the lowest intertidal heights face **competition and predation** from mobile invertebrates (low boundary), while at the highest heights they face **desiccation and thermal stress** (upper boundary). A quadratic model is the minimal polynomial that can capture this unimodal distribution.

c.

- ✓ Method 1: **Nested F-test** comparing the linear and quadratic models using `anova(lm_linear, lm_quadratic)` in R. This directly tests whether the quadratic term explains significantly more variance than expected by chance: $F = (SS_{\sim\text{quadratic term}}) / (MS_{\text{residuals,quadratic}})$. A significant $F (p < 0.05)$ favours the quadratic model.
- ✓ Method 2: **AIC comparison**: compute AIC for both models; the model with the lower AIC is preferred. A $\Delta\text{AIC} > 2$ constitutes meaningful evidence in favour of the quadratic model; $\Delta\text{AIC} > 10$ constitutes strong evidence.

d.

- ✓ A high-degree polynomial (e.g., degree 4 or 5) can produce wild oscillations between data points and extreme extrapolations beyond the observed height range – curves that pass through every data point but have no biological meaning.
- ✓ Ecologically, such a model might suggest multiple “peaks” or “valleys” in barnacle density at specific intertidal heights that are artefacts of sample noise, not real gradients. Predictions beyond the sampled range would be severely biased. The goal is the **simplest model** (fewest parameters) that adequately describes the true biological pattern – a quadratic is sufficient here, and adding higher-degree terms risks fitting noise rather than signal.

Part C: Statistical Output Interpretation (37 marks)

Question 13 – Wilcoxon Signed-Rank Test Output (/12)

A respiratory physiologist measures oxygen consumption ($\text{mL O}_2 \text{ kg}^{-1} \text{ min}^{-1}$) in 18 individual yellowtail kingfish (*Seriola lalandi*) at rest and immediately after a standardised burst-swimming event. The same fish are measured at both time points. The data failed the Shapiro-Wilk test for normality of differences. The output is:

```
Wilcoxon signed-rank exact test

data: oxygen_mL_kg by time_point
V = 162, p-value = 0.000214
alternative hypothesis: true location shift is not equal to 0

95 percent confidence interval:
 9.84 27.31
sample estimates:
(pseudo)median
 18.43
```

- State the null hypothesis tested by this analysis. (/ 2)
- What does the test statistic $V = 162$ represent conceptually? (/ 2)
- Interpret the p -value = 0.000214 at $\alpha = 0.05$. What conclusion do you draw? (/ 3)
- Interpret the 95% confidence interval (9.84, 27.31) and the pseudomedian (18.43) in biological terms. (/ 3)
- Why was the Wilcoxon signed-rank test used rather than a paired t -test? (/ 2)

💡 Tip

Model Answer – Question 13

a.

- ✓ H_0 : There is no difference in oxygen consumption between rest and post-burst-swimming – the median of the paired differences (post – rest) equals zero; the location shift is 0.
- ✓ (Equivalently: the distribution of paired differences is symmetric around zero – the Wilcoxon signed-rank test's null.)

b.

- ✓ V is the **sum of the positive signed ranks** – after computing the absolute differences between paired measurements, ranking them from smallest to largest, and assigning positive signs to ranks where the post-swimming value exceeds the resting value, V is the total of those positive ranks.
- ✓ Under H_0 (no location shift), positive and negative ranks would be equally likely and V would equal approximately half the maximum possible sum. A large V relative to the null expectation indicates that post-swimming oxygen consumption is systematically higher than resting consumption.

c.

- ✓ $p = 0.000214 \ll \alpha = 0.05$: we **strongly reject** H_0 . There is overwhelming statistical evidence that oxygen consumption increases significantly following burst swimming in yellowtail kingfish.
- ✓ The probability of observing a signed-rank statistic as extreme as $V = 162$ by chance, if there were truly no difference between rest and post-exercise oxygen consumption, is only 0.021% – an extremely unlikely result under the null hypothesis.

d.

- ✓ The **pseudomedian** ($18.43 \text{ mL O}_2 \text{ kg}^{-1} \text{ min}^{-1}$) is the estimated median of the paired differences – the typical (median) increase in oxygen consumption following burst swimming is approximately **$18.4 \text{ mL O}_2 \text{ kg}^{-1} \text{ min}^{-1}$** .
- ✓ The **95% CI (9.84, 27.31)** means we are 95% confident that the true median increase in oxygen consumption lies between approximately **9.8 and $27.3 \text{ mL O}_2 \text{ kg}^{-1} \text{ min}^{-1}$** . Because this interval excludes zero, it independently confirms statistical significance. The interval is relatively wide, reflecting uncertainty about the precise magnitude of the exercise-induced oxygen debt.

e.

- ✓ The Wilcoxon signed-rank test was used because the **paired differences failed the Shapiro-Wilk normality test** – a prerequisite for the paired t -test (which assumes the paired differences are approximately normally distributed).
- ✓ The Wilcoxon signed-rank test is a **non-parametric alternative that operates on the ranks of the absolute differences rather than the raw values**, requiring only that the differences are symmetrically distributed – a weaker and more easily satisfied assumption when sample sizes are modest ($n = 18$).

Question 14 – Two-Way ANOVA Output: Prawn Survival (/12)

An aquaculturist tests the effects of **salinity** (three levels: 20, 30, 40 ppt) and **temperature** (two levels: 20°C, 28°C) on the percentage survival (%) of juvenile prawns (*Penaeus japonicus*) after 72 hours. Ten replicates per treatment combination are used. The ANOVA table is:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
salinity	2	1284.6	642.3	31.48	< 0.001 ***
temperature	1	394.1	394.1	19.32	< 0.001 ***
salinity:temperature	2	187.3	93.7	4.59	0.0132 *
Residuals	54	1101.8	20.4		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- State the null hypothesis for the **salinity × temperature interaction** term. (/ 2)
- What does **Residuals df = 54** tell you about the experimental design? Show your reasoning. (/ 2)
- The interaction is significant ($F(2, 54) = 4.59, p = 0.0132$). Interpret this biologically. (/ 3)
- Because the interaction is significant, how should you approach the interpretation of the main effects? (/ 3)
- Verify the F -value for salinity using the values in the ANOVA table. Show your working. (/ 2)

 Tip

Model Answer – Question 14

a.

- ✓ H_0 : The effect of salinity on juvenile prawn survival is the same at both temperatures – the two factors act **additively**, and the response to salinity does not depend on temperature (or equivalently, the temperature effect is the same at all three salinities).

b.

- ✓ Total observations n = number of cells × replicates per cell = (3 salinity × 2 temperature) × 10 = **60 observations**.
- ✓ Residuals df = total n – number of cells = 60 – 6 = **54** ✓. This represents the within-cell error – the variability among the 10 replicates within each salinity × temperature treatment combination.

c.

- ✓ A significant interaction means the **effect of salinity on survival depends on temperature** – the response to changing salinity is not the same at 20°C as it is at 28°C.
- ✓ Biologically: at 20°C, prawns may tolerate a broader range of salinities (a flat, moderate survival plateau), while at 28°C (thermal stress), deviations from optimal salinity (30 ppt) may be far more lethal – the additional stressor of temperature reduces the physiological scope to cope with osmotic imbalance.
- ✓ This non-additive (synergistic stressor) interaction has important practical implications: the optimal salinity for prawn aquaculture depends on the culture temperature and cannot be determined from single-factor experiments.

d.

- ✓ When the interaction is significant, the main effects are **conditional** – the effect of salinity cannot be summarised as a single universal value, because it differs between the two temperature levels. Reporting “salinity significantly affected survival” without qualification is misleading.
- ✓ The appropriate approach is to interpret and present **simple main effects** – the effect of salinity separately at 20°C and at 28°C, ideally with an interaction plot. Main effects should not be interpreted in isolation when a significant interaction is present.

e.

- ✓ $F_{\text{salinity}} = \text{Mean Square}_{\text{salinity}} / \text{Mean Square}_{\text{Residuals}} = 642.3 / 20.4 = 31.48$ ✓
- ✓ This F -ratio means the between-salinity-group variance is 31.48 times larger than the within-cell residual variance – the salinity differences are far beyond what would be expected from random sampling of a common population, strongly supporting rejection of the salinity H_0 .

Question 15 – Multiple Regression Output: Bird Territory Size (/13)

An ornithologist models breeding territory size (ha) of *Acrocephalus scirpaceus* (reed warbler) as a function of three continuous environmental predictors: mean vegetation height (m), local prey density (invertebrates m^{-2}), and conspecific density (breeding pairs km^{-2}). The `lm()` output, with VIF values, is:

```
Call:
lm(formula = territory_ha ~ veg_height_m + prey_density + conspecific_density,
    data = warblers)

Residuals:
    Min       1Q   Median       3Q      Max
-7.812  -2.134   0.218   2.019   8.431

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.321     0.812    5.32 < 0.001 ***
veg_height_m      2.834     0.381    7.44 < 0.001 ***
prey_density     -1.412     0.293   -4.82 < 0.001 ***
conspecific_density -0.887     0.412   -2.15  0.0340 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.214 on 81 degrees of freedom
Multiple R-squared:  0.6812, Adjusted R-squared:  0.6697
F-statistic: 57.91 on 3 and 81 DF, p-value: < 2.2e-16

Variance Inflation Factors:
              veg_height_m      prey_density conspecific_density
                1.83                2.14                2.41
```

- Write the fitted regression equation. (/ 2)
- Interpret each of the three slope coefficients in **biological terms**, specifically using the word “partial.” (/ 4)
- Examine the VIF values. What do they suggest about multicollinearity in this model? Does the model have a collinearity problem? (/ 3)
- What does $R^2 = 0.6812$ indicate about this model? (/ 2)
- What does the overall F -statistic ($F(3, 81) = 57.91$) test, and what does the significant p -value tell you? (/ 2)

💡 Tip

Model Answer – Question 15

a.

• ✓ $\widehat{territory_ha} = 4.321 + 2.834 \times veg_height_m - 1.412 \times prey_density - 0.887 \times conspecific_density$

b. (Accept partial credit for each coefficient)

- ✓ **veg_height_m (2.834)**: each 1 m increase in vegetation height is associated with a **partial** increase of approximately 2.83 ha in territory size, holding prey density and conspecific density constant. Taller reed beds may support larger territories because they provide more concealment, nest-site options, and foraging microhabitats.
- ✓ **prey_density (-1.412)**: each additional invertebrate m^{-2} of prey density is associated with a **partial** decrease of approximately 1.41 ha in territory size, holding vegetation height and conspecific density constant. Birds can maintain smaller territories when food is more concentrated – territory size is inversely related to prey availability.
- ✓ **conspecific_density (-0.887)**: each additional breeding pair km^{-2} of conspecific density is associated with a **partial** decrease of approximately 0.89 ha in territory size, holding the other predictors constant. Higher conspecific pressure likely forces territory compression through competitive exclusion.

c.

- ✓ All three VIF values are low: 1.83, 2.14, and 2.41 – all well below the conventional threshold of 5 (or 10). This indicates **negligible multicollinearity** – the three predictors are only weakly correlated with each other and share little variance in common.
- ✓ The model does **not** have a collinearity problem. The partial regression coefficients are estimated stably (small standard errors relative to estimates), and each predictor's contribution can be interpreted independently.

d.

- ✓ $R^2 = 0.6812$ means that **68.1% of the total variation in reed warbler territory size** is explained by vegetation height, prey density, and conspecific density together. This is a moderate-to-strong fit – the three environmental predictors account for more than two-thirds of the variation in territory size across breeding sites.
- ✓ The remaining ~32% of variation is attributable to unmeasured factors (e.g., within-site nest-site quality, individual territory-holder quality, prior residency effects).

e.

- ✓ The overall F -test evaluates whether the **model as a whole** – using all three predictors simultaneously – explains significantly more variance than the intercept-only null model (i.e., tests $H_0: \beta_1 = \beta_2 = \beta_3 = 0$).
- ✓ $F(3, 81) = 57.91, p < 2.2 \times 10^{-16}$: we strongly reject the null model. The combination of vegetation height, prey density, and conspecific density provides a significantly better description of territory size than the hypothesis that none of these variables matters. At least one predictor (and in this case all three are individually significant) is a real predictor of territory size.

End of Version 9

Bibliography