BCB744 Biostatistics — Theory Test (Version 6)

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 — Repeated Measures and Within-Subject Designs (/7)

a. Explain the difference between a **between-subjects** and a **within-subjects (repeated measures)** experimental design. **(/ 2)**

b. What statistical advantage does a within-subjects design offer over a between-subjects design, and why? **(/ 2)**

c. A researcher measures plant biomass at weeks 0, 4, 8, and 12 under two fertiliser treatments (control, high-N). The same 15 pots are measured at all four time points. What type of analysis is most appropriate, and what violation of standard one-way ANOVA assumptions must be addressed? **(/ 3)**

Tip

Model Answer — Question 1

a.

- ✓ In a **between-subjects design**, different individuals are assigned to different treatment conditions — each person or experimental unit contributes data to only one group. Variation between individuals is part of the error term.
- ✓ In a **within-subjects (repeated measures) design**, the same individuals are measured under all treatment conditions or at multiple time points — each unit contributes observations to every level of the within-subjects factor. The design exploits the fact that each individual serves as its own control.

b.

- ✓ A within-subjects design **reduces residual variance** and thereby **increases statistical power**, because individual-level baseline differences (which are a major source of noise) are removed from the error term by subtracting each subject's mean response.
- ✓ Mechanistically: in a between-subjects design, between-individual variability is part of the error; in a within-subjects design, this variability is partitioned into a separate subject term and is excluded from the denominator of the F-ratio, leaving only within-subject variability as error.

c.

- ✓ The appropriate analysis is a **two-way repeated measures ANOVA** (or a linear mixed-effects model), with time as the within-subjects factor and fertiliser treatment as the between-subjects factor.
- ✓ Standard ANOVA assumes that observations are **independent**, but repeated measurements from the same pot are correlated — this violates the independence assumption. Additionally, repeated measures ANOVA requires **sphericity**: the variances of the differences between all pairs of time points must be equal. This is checked with **Mauchly's test of sphericity**; if violated, epsilon corrections (Greenhouse-Geisser or Huynh-Feldt) are applied to the degrees of freedom.

Question 2 — A Priori Power Analysis (/6)

a. What is an **a priori power analysis**, and at what stage of a study should it be conducted? **(/ 2)**

b. List the **four quantities** that determine statistical power. Explain briefly how they are interrelated. **(/ 2)**

c. A researcher plans an independent-samples t-test and needs to detect a medium effect size (Cohen's $d = 0.5$) with power = 0.80 at $\alpha = 0.05$. A power calculation yields $n = 64$ per group. How does the required n change if (i) they increase the required power to 0.90, and (ii) they decide the minimum detectable effect is $d = 0.3$ instead? **(/ 2)**

Tip

Model Answer — Question 2

a.

- ✓ An **a priori power analysis** calculates the minimum sample size required to detect an effect of a specified magnitude with a given probability, before data are collected. It is conducted at the **study design stage** — prior to any data collection — to ensure the study is adequately powered to answer its research question.
- ✓ Conducting it after data collection (post-hoc power analysis based on observed effects) is uninformative and potentially misleading because observed power is tautologically determined by the obtained p-value.

b.

- ✓ The four quantities: (1) **significance level** (α, Type I error rate), (2) **statistical power** (1 − β, where β is the Type II error rate), (3) **effect size** (the magnitude of the true biological difference relative to variability), and (4) **sample size** (n).
- ✓ They are interrelated: for fixed α and effect size, increasing n increases power; for fixed n and α, a larger effect size is easier to detect (higher power); for fixed n and effect size, reducing α (making the test more stringent) reduces power (increases β).

c.

- ✓ (i) Increasing required power from 0.80 to 0.90 requires a **larger** sample size (approximately $n \approx 85$ per group instead of 64) — higher power means a lower acceptable Type II error, which demands more data to reliably detect the effect.
- ✓ (ii) Reducing the minimum detectable effect from $d = 0.5$ to $d = 0.3$ requires a **much larger** sample size (approximately $n \approx 176$ per group) — smaller effects are harder to detect and require more observations to distinguish from random noise.

Question 3 — Influential Observations and Cook's Distance (/7)

a. Distinguish between an **outlier** and an **influential observation** in the context of regression. Can a data point be one without being the other? **(/ 3)**

b. What is **Cook's distance**, and what does a large value indicate? What threshold is commonly used to identify potentially problematic points? **(/ 2)**

c. A researcher finds one observation with Cook's distance = 1.8 after fitting a simple linear regression. Describe the steps they should take before deciding whether to remove it. **(/ 2)**

Tip

Model Answer — Question 3

a.

- ✓ An **outlier** is a data point with an unusually large residual — its observed y value deviates substantially from the model's prediction. An outlier has high **residual leverage** in the y-direction.
- ✓ An **influential observation** is one that, if removed, would substantially change the fitted model (e.g., coefficients, slope, intercept). A point is influential if it has high **leverage** (an extreme x value, far from the mean of the predictor) or a large residual, or both.
- ✓ Yes, a point can be one without the other: a data point with extreme x but a y value that falls exactly on the regression line has high leverage but is not an outlier (it may even pull the line toward itself, making it appear to fit well). Conversely, an outlier near the mean of x has a large residual but little influence on the slope.

b.

- ✓ **Cook's distance** for observation i measures the aggregate change in all fitted values when observation i is deleted from the analysis — it combines the residual size and leverage into a single influence measure.
- ✓ A large Cook's distance (typically **> 1** is flagged as potentially problematic; some use $4/n$ as a threshold for large samples) indicates that removing this observation would substantially alter the fitted model. Such points warrant careful investigation.

c.

- ✓ Step 1: **Examine the raw data** for observation i — check for data entry errors, transcription mistakes, or equipment malfunctions. A recording error is a valid reason for exclusion; genuine biological extreme values are not.
- ✓ Step 2: **Fit the model with and without** the observation and compare the key results (slope estimate, standard error, p-value, R^2). If conclusions change meaningfully, the point is genuinely influential; report sensitivity analysis results transparently.
- ✓ The researcher should **not remove the point** solely because it is influential — only if there is a verified non-biological reason for its extreme value, or if both the "with" and "without" models are reported to show the robustness (or fragility) of the conclusion.

Question 4 — Standardised Regression Coefficients (/6)

a. What is a **standardised (beta) regression coefficient**, and how does it differ from an unstandardised coefficient? **(/ 3)**

b. In a multiple regression predicting bird species richness from habitat patch area (ha) and distance to the nearest forest fragment (km), the standardised coefficients are $\beta_{\text{area}} = 0.61$ and $\beta_{\text{distance}} = -0.38$. What can you conclude about the **relative importance** of the two predictors? **(/ 2)**

c. Why can standardised coefficients **not** be meaningfully compared across different studies that used different samples? **(/ 1)**

Tip

Model Answer — Question 4

a.

- ✓ An **unstandardised coefficient** (b) gives the change in the response variable (in its original units) for a one-unit increase in the predictor (in its original units). Its value depends on the measurement scales of both variables, making direct comparison of coefficients across predictors (or studies) with different units meaningless.
- ✓ A **standardised (beta) coefficient** is obtained by standardising both the response and predictor variables to have mean = 0 and SD = 1 before fitting the model (or equivalently, by multiplying the unstandardised coefficient by SD_x / SD_y). It gives the change in the response in standard deviation units for a one-SD increase in the predictor.
- ✓ Standardised coefficients are **unitless** and can be compared directly across predictors within the same model, providing a measure of the relative contribution of each predictor.

b.

- ✓ Patch area ($\beta = 0.61$) has a **larger absolute standardised coefficient** than distance to fragment ($\beta = -0.38$), indicating that a one-SD increase in patch area is associated with a larger change in species richness (in SD units) than a one-SD increase in distance.
- ✓ Therefore, within this model and dataset, **patch area is the more important predictor** of bird species richness. Distance has a moderate negative effect (larger distance → fewer species), but patch area explains more of the variation.

c.

- ✓ Standardised coefficients are scaled by the **standard deviation of the predictor** in the sample used. If two studies sample populations with different ranges or variances of the predictor (e.g., one study covers a small patch-size range, another a wide range), the standard deviations will differ, and the same underlying unstandardised slope will produce different beta weights. Comparing beta coefficients across studies therefore conflates the underlying effect size with sample variability.

Question 5 — Poisson Generalised Linear Models (/8)

a. What is the **canonical link function** for a Poisson GLM, and why is it used? What constraint on the response variable makes the Poisson family appropriate? **(/ 3)**

b. A Poisson GLM of insect species richness (count) on vegetation cover (%) returns a coefficient of $\beta = 0.021$ for cover. Interpret this coefficient on the **response scale** (not the log scale). **(/ 3)**

c. How is the **overall significance** of a Poisson GLM assessed? What statistic and distribution are used? **(/ 2)**

Tip

Model Answer — Question 5

a.

- ✓ The canonical (default) link function for the Poisson family is the **natural logarithm** (log link): $\log(\mu) = \eta$, where μ is the expected count and η is the linear predictor. The inverse link (exponential) maps the linear predictor back to the expected count scale, ensuring predicted counts are always **positive** — a necessary constraint for count data.
- ✓ The Poisson family is appropriate when the response is **non-negative integer-valued count data** representing the number of events in a fixed interval or area (e.g., number of species per plot, number of eggs per nest), and when the mean and variance of the counts are approximately equal (equidispersion).

b.

- ✓ On the log scale, the coefficient $\beta = 0.021$ means that $\log(\mu)$ increases by 0.021 for each 1% increase in vegetation cover. On the **response (count) scale**, the expected count is multiplied by $e^{0.021} \approx$ **1.021** for each 1% increase in cover — a 2.1% increase in expected insect species richness per 1% increase in cover.
- ✓ For a 10% increase in cover, the multiplier is e^(0.021 × 10)^ = $e^{0.21} \approx 1.23$ — a 23% increase in expected species count. Biologically: greater vegetation cover is associated with higher insect diversity, consistent with habitat complexity and resource diversity hypotheses.

c.

- ✓ Overall significance is assessed via a **likelihood ratio test** comparing the fitted model's deviance to the null model's deviance (intercept only). The test statistic is Δdeviance = null deviance − residual deviance, which follows a χ^2 **distribution** with degrees of freedom equal to the number of added parameters.
- ✓ Alternatively, individual coefficients are tested using **Wald z-tests** (Estimate / SE, compared to the standard normal distribution), as shown in the model output.

Question 6 — Sampling Strategies (/7)

a. Distinguish between **simple random sampling**, **stratified random sampling**, and **systematic sampling**. For each, give one ecological situation where it would be the preferred approach. **(/ 4)**

b. Why does **convenience sampling** (measuring whatever is easiest to access) undermine the validity of statistical inference, even when the sample size is large? **(/ 2)**

c. What is **cluster sampling**, and in what field situation might it be necessary? **(/ 1)**

Tip

Model Answer — Question 6

a.

- ✓ **Simple random sampling**: every individual in the population has an equal and independent probability of being selected. Preferred when the population is relatively homogeneous and accessible — e.g., randomly selecting individual fish from a well-mixed hatchery tank for growth measurements.
- ✓ **Stratified random sampling**: the population is divided into subgroups (strata) based on a known characteristic, and random samples are drawn from each stratum. Preferred when the population is heterogeneous and strata are identifiable — e.g., sampling intertidal invertebrates separately from low, mid, and high tidal zones to ensure all zones are represented.
- ✓ **Systematic sampling**: individuals are selected at regular intervals from a list or spatial transect (e.g., every 10th individual, every 5 m along a transect). Preferred in large, spatially continuous surveys where a random start and fixed interval gives practical coverage — e.g., sampling vegetation along a belt transect at fixed 5-m intervals to characterise plant community composition.

b.

- ✓ Convenience sampling creates a **biased sample** — the accessible individuals are systematically different from the population as a whole (e.g., rocky intertidal species sampled only near a car park may differ from those on remote shores in exposure, disturbance, or size). Statistical inference assumes a representative sample; with a biased sample, the estimated parameters (means, proportions, regression slopes) reflect the accessible subpopulation, not the target population.
- ✓ A large biased sample does not reduce this bias — it simply gives a more precise estimate of the wrong quantity (a more precise answer to the wrong question).

c.

- ✓ **Cluster sampling** divides the population into naturally occurring groups (clusters) and randomly samples entire clusters, then measures all (or a random subset of) individuals within selected clusters. It is used when it is impractical to list all individuals in advance or when individuals are naturally grouped — e.g., surveying all trees within randomly selected 1-ha forest plots, rather than selecting individual trees at random across a vast landscape.

Question 7 — Bootstrap and Permutation Methods (/7)

a. What is a **bootstrap confidence interval**? Describe the general procedure in plain language. **(/ 3)**

b. How does a **permutation test** differ from a parametric hypothesis test in terms of how the null distribution is constructed? **(/ 2)**

c. Give **two specific situations** in which a bootstrap or permutation method would be preferred over a standard parametric test. **(/ 2)**

Tip

Model Answer — Question 7

a.

- ✓ A **bootstrap confidence interval** is constructed by repeatedly resampling from the observed data **with replacement** to approximate the sampling distribution of the statistic of interest — without assuming a particular parametric distribution.
- ✓ Procedure: (1) Draw B bootstrap samples (e.g., 10,000), each of size n with replacement from the original data. (2) Calculate the statistic of interest (e.g., the mean, median, or regression coefficient) for each bootstrap sample, obtaining a distribution of B bootstrap estimates. (3) Use the 2.5th and 97.5th percentiles of this distribution as the lower and upper bounds of the 95% bootstrap CI.
- ✓ The bootstrap treats the sample as a proxy for the population, using its own distribution to approximate variability rather than relying on theoretical assumptions such as normality.

b.

- ✓ A **parametric hypothesis test** constructs the null distribution using mathematical theory (e.g., the t-distribution, F-distribution, or normal approximation) derived from distributional assumptions about the data.
- ✓ A **permutation test** constructs the null distribution **empirically from the data** by randomly shuffling (permuting) the observed values between groups many times and computing the test statistic each time. The null distribution reflects what would be expected if there were no relationship between the predictor and response — no distributional assumptions are needed.

c.

- ✓ Any two valid situations: (1) **small sample sizes** where parametric distributional assumptions (e.g., normality) cannot be verified — bootstrap/permutation methods remain valid without these assumptions; (2) **non-standard statistics** for which no analytical sampling distribution exists (e.g., confidence intervals for a ratio of medians, or for Cronbach's α) — bootstrap directly approximates the sampling distribution of any computable statistic; (3) heavily skewed, multimodal, or otherwise irregularly distributed data where parametric tests are inappropriate even with moderate sample sizes.

Question 8 — Generalised Linear Models — Framework (/6)

a. What are the **three components** that define a generalised linear model (GLM)? Describe each briefly. **(/ 3)**

b. For each of the following response variables, identify the most appropriate **GLM family and link function**: (i) number of nesting attempts per breeding season (count, non-negative

integer); (ii) proportion of larvae surviving to metamorphosis (bounded 0–1, based on a known denominator). **(/ 2)**

c. Why is it incorrect to apply a standard Gaussian (normal-errors) GLM to a response variable that consists of counts of rare events (many zeros, right-skewed)? **(/ 1)**

Tip

Model Answer — Question 8

a.

- ✓ **Random component**: specifies the probability distribution of the response variable — the error distribution (e.g., Poisson for counts, binomial for proportions, Gaussian for continuous unbounded data). It determines how variance scales with the mean.
- ✓ **Systematic component**: the linear predictor $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ — a linear combination of the predictor variables and their coefficients, identical in form to ordinary linear regression.
- ✓ **Link function**: the mathematical function $g(\mu) = \eta$ that connects the expected value of the response (μ) to the linear predictor. It transforms the mean to an unbounded scale so that the linear predictor can model it without constraint violations.

b.

- ✓ (i) Nesting attempts (count data): **Poisson family with log link**. Counts are non-negative integers; the log link ensures predicted values remain positive; Poisson distribution is the natural choice for counts of independent events.
- ✓ (ii) Proportion surviving (bounded 0–1 with known denominator): **Binomial family with logit link**. The response is a proportion arising from a fixed number of trials (larvae); the logit link maps (0, 1) to $(-\infty, +\infty)$; the binomial distribution models the number of successes from n trials.

c.

- ✓ A Gaussian GLM (normal errors) assumes an unbounded, continuous, symmetric response with constant variance. Count data that are mostly zeros are **non-negative**, **discrete**, **right-skewed**, and typically show variance that increases with the mean — all properties incompatible with Gaussian assumptions. Applying a Gaussian model would produce negative predicted counts (impossible), severely violate the normality and homoscedasticity assumptions, and give invalid standard errors and p-values.

Question 9 — Nested Experimental Designs (/7)

a. What is a **nested (hierarchical) experimental design**? Give a biological example with at least three levels of nesting. **(/ 3)**

b. Why is it statistically incorrect to treat **sub-samples** (e.g., multiple tissue sections from the same organism) as independent replicates in a nested design? **(/ 2)**

c. What is the **correct experimental unit** for testing a treatment effect when treatments are applied at one level but measurements are made at a lower level? **(/ 2)**

Tip

Model Answer — Question 9

a.

- ✓ A **nested design** is one in which observations at lower levels of organisation are grouped within higher-level units, with each lower-level unit belonging to exactly one higher-level unit. The levels are not crossed — each combination appears within only one higher-level group (unlike factorial designs).
- ✓ **Example**: measuring liver enzyme activity in fish from multiple rivers, with multiple fish per river and multiple tissue samples per fish. The three levels are: River (broadest) → Fish (nested within river) → Tissue sample (nested within fish). Treatment (e.g., pollution level) is applied at the river level; measurements are made on individual tissue samples.

b.

- ✓ Sub-samples from the same organism share a common biological environment — same genetics, diet, physiology, and treatment history. They are therefore **not independent**: the values are more similar to each other (correlated within the organism) than to values from a different organism, even under the same treatment.
- ✓ Treating sub-samples as independent replicates inflates the apparent n, underestimates the within-treatment variance, and dramatically inflates the test statistic and Type I error rate — this is **pseudoreplication**.

c.

- ✓ The correct experimental unit is the **unit at the level to which the treatment is applied** — because that is the unit that receives an independent assignment. In the fish example, the **river** (or the individual fish, if treatment is applied per fish) is the experimental unit for testing the treatment effect. The tissue samples are sub-samples (pseudo-replicates) and should be averaged within each experimental unit before analysis, or modelled using a nested ANOVA or mixed-effects model.

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

Question 10 — Repeated Measures: Trout Immune Response (/13)

A fish immunologist measures plasma antibody titre (arbitrary units) of 18 individual rainbow trout (*Oncorhynchus mykiss*) at three time points: Baseline (Day 0), post-vaccination (Day 14), and

peak immune response (Day 28). The same 18 fish are measured at all three time points. The first six rows of the dataset are:

```
  fish_id  timepoint  antibody_titre
1       1        D0             1.2
2       1       D14             4.8
3       1       D28             9.1
4       2        D0             1.4
5       2       D14             5.2
6       2       D28             8.7
```

The research question is: *"Does plasma antibody titre change significantly over time following vaccination?"*

a. State formal null and alternative hypotheses for this analysis. **(/ 3)**
b. Identify the appropriate statistical test and give **three reasons** for your choice, with reference to the study design and data structure. **(/ 5)**
c. What key assumption does this repeated measures design introduce that standard one-way ANOVA does not require? How is it checked? **(/ 3)**
d. If the overall test is significant, what **post-hoc procedure** would you apply to identify which time points differ? **(/ 2)**

Tip

Model Answer — Question 10

a.

- ✓ H_0: Mean plasma antibody titre does not differ among time points; $\mu_{D0} = \mu_{D14} = \mu_{D28}$ (there is no change in antibody titre over time following vaccination).
- ✓ H_A: Mean antibody titre differs among at least some time points — vaccination produces a significant change in immune titre over time.
- ✓ The alternative is non-directional for the omnibus test (though a directional prediction of increasing titre is biologically motivated, the overall *F*-test is non-directional; directional predictions should be addressed in post-hoc comparisons).

b.

- ✓ **One-way repeated measures ANOVA** (or a linear mixed-effects model with time as a fixed effect and fish_id as a random intercept).
- ✓ Reason 1: The same **18 individual fish** are measured at all three time points — the measurements are not independent between time points within the same fish. This within-subject structure requires a repeated measures approach.
- ✓ Reason 2: There is a **single categorical within-subjects factor** (timepoint) with **three levels** (D0, D14, D28). A one-way repeated measures ANOVA is designed to test for differences in a continuous response across multiple levels of a within-subjects factor. Using three paired *t*-tests would inflate the Type I error rate.
- ✓ Reason 3: The **response variable** (antibody titre) is continuous, meeting the measurement-scale requirement for a parametric approach.

c.

- ✓ The key additional assumption is **sphericity**: the variances of the differences between all pairs of time points must be equal — i.e., Var(D0 − D14) = Var(D0 − D28) = Var(D14 − D28).
- ✓ Sphericity is checked using **Mauchly's test of sphericity**. If violated (Mauchly's $p < 0.05$), the *F*-statistic is positively biased, and degrees of freedom must be corrected using the Greenhouse-Geisser or Huynh-Feldt epsilon correction to maintain a valid α level.

d.

- ✓ **Pairwise paired *t*-tests with a Bonferroni (or Holm) correction** are applied to compare all pairs of time points (D0 vs. D14, D0 vs. D28, D14 vs. D28), correcting the *p*-values for the three simultaneous comparisons.
- ✓ Alternatively, **Tukey's HSD** (if available in the repeated-measures framework) or a linear contrast approach can be used to identify the specific time points at which significant changes in titre occur.

Question 11 — Multiple Regression: Seagrass Biomass (/12)

A coastal ecologist measures above-ground seagrass biomass (g m^{-2}) at 48 sites, along with three candidate environmental predictors: water clarity (Secchi depth, m), tidal exposure (hours exposed per day), and sediment organic matter (%). The first six rows of the dataset are:

	site	biomass_g_m2	secchi_m	tidal_hrs	sediment_om_pct
1	1	412.3	3.2	2.1	3.4
2	2	387.1	2.9	2.8	4.1
3	3	318.5	2.1	3.5	5.2
4	4	271.4	1.8	4.2	6.8
5	5	224.8	1.4	5.1	8.3
6	6	193.2	1.2	5.8	9.1

The research aim is: *"To determine which combination of environmental variables best predicts above-ground seagrass biomass."*

a. State the null and alternative hypotheses for the **overall** multiple regression model. **(/ 3)**
b. Give **three specific reasons** why multiple regression (not simple linear regression) is appropriate for this research aim. **(/ 3)**
c. The researcher uses AIC to compare four candidate models. Describe what they would look for in the AIC table to select the best model. **(/ 3)**
d. From the data preview, what concern arises about **multicollinearity** among the three predictors, and how would you formally diagnose it? **(/ 3)**

Tip

Model Answer — Question 11

a.

- ✓ H_0: None of the three environmental predictors (Secchi depth, tidal exposure, sediment OM) has a linear relationship with seagrass biomass; all regression slopes ($\beta_1 = \beta_2 = \beta_3$) = 0. The model explains no more variance than the intercept-only model.
- ✓ H_A: At least one predictor has a non-zero slope — at least one environmental variable is a significant linear predictor of seagrass biomass.
- ✓ The overall null is tested by the omnibus *F*-statistic in the ANOVA table of the regression output.

b.

- ✓ Reason 1: There are **three candidate predictors**, each potentially influencing biomass. Simple linear regression models only one predictor at a time, ignoring the simultaneous effects of the others and failing to control for their confounding influence on the focal predictor's coefficient.
- ✓ Reason 2: The research aim is to identify the **best combination** of predictors — this requires fitting and comparing models that include subsets of all three predictors simultaneously, which is the purpose of multiple regression and model selection.
- ✓ Reason 3: **Partial regression coefficients** in multiple regression estimate the effect of each predictor **holding the others constant** — providing a clearer picture of each variable's independent contribution. Simple regression slopes confound the effects of correlated predictors.

c.

- ✓ Select the model with the **lowest AIC value** as the best supported model. Compare other candidate models by computing $\Delta AIC = AIC_{model} - AIC_{minimum}$. Models with $\Delta AIC < 2$ are considered empirically equivalent (equally well supported); $\Delta AIC > 10$ indicates strong evidence against the model.
- ✓ Also check that the "best" model makes biological sense — AIC minimisation should be combined with substantive knowledge. A simpler model (fewer parameters) with nearly equal AIC may be preferred on the principle of parsimony (prefer AICc for small n/k ratios).

d.

- ✓ The data preview shows that Secchi depth decreases, tidal exposure increases, and sediment OM increases together as site number increases — the three predictors appear to **co-vary systematically**, suggesting strong inter-predictor correlations (multicollinearity). Sites with more tidal exposure may have lower water clarity and higher organic matter deposition.
- ✓ Formal diagnosis: calculate the **Variance Inflation Factor (VIF)** for each predictor after fitting the full model. $VIF_j = 1 / (1 - R^2_j)$, where R^2_j is the variance in predictor *j* explained by all other predictors. VIF > 5 (or > 10 by more lenient standards) indicates problematic multicollinearity that inflates coefficient standard errors and makes individual estimates unstable.

Question 12 — Randomised Complete Block Design: Kelp Fertilisation (/12)

A marine botanist tests the effect of three nutrient treatments (Control, Low-N, High-N) on the daily growth rate (mm day^{-1}) of juvenile kelp (*Ecklonia maxima*). The experiment is conducted in six seawater flow-through tanks (blocks); within each tank, three kelp juveniles are randomly assigned one treatment each. The first nine rows of the data are:

```
  tank  treatment  growth_mm_d
1    1    Control          3.2
2    1      Low-N          5.1
3    1     High-N          8.4
4    2    Control          2.9
5    2      Low-N          4.8
6    2     High-N          7.9
7    3    Control          3.4
8    3      Low-N          5.5
9    3     High-N          8.8
```

The research question is: *"Does nutrient treatment significantly affect daily growth rate of juvenile kelp?"*

a. State the formal null and alternative hypotheses for the treatment effect. **(/ 2)**
b. Identify the most appropriate statistical test and give **three reasons** for your choice, with specific reference to the experimental design. **(/ 5)**
c. What is the purpose of including `tank` as a blocking factor in the analysis? What variation does it account for? **(/ 3)**
d. If you ignored the blocking factor and ran a simple one-way ANOVA instead, how would this affect the test's ability to detect treatment differences? **(/ 2)**

Tip

Model Answer — Question 12

a.

- ✓ H_0: Mean daily kelp growth rate does not differ among the three nutrient treatments ($\mu_{Control} = \mu_{Low\text{-}N} = \mu_{High\text{-}N}$).
- ✓ H_A: At least one nutrient treatment produces a mean growth rate that differs significantly from the others.

b.

- ✓ **One-way ANOVA with blocks** — a **randomised complete block design (RCBD) ANOVA** (`lm(growth_mm_d ~ treatment + tank)` in R), or equivalently a two-way ANOVA treating tank as a second (blocking) factor.
- ✓ Reason 1: The response variable (growth rate, mm day^{-1}) is **continuous** and the predictor of interest (treatment) is **categorical with three levels** — the design is directly analogous to a one-way ANOVA. Multiple pairwise *t*-tests would inflate the family-wise error rate.
- ✓ Reason 2: The design is a **randomised complete block design**: each tank contains all three treatment levels, making tank a blocking factor that must be included in the model. Ignoring blocking violates the independence structure of the design.
- ✓ Reason 3: Including the tank (block) term in the ANOVA model **accounts for between-tank variation** in baseline growth conditions (e.g., differences in water flow, temperature, light availability among tanks), reducing the residual variance and increasing sensitivity to treatment differences.

c.

- ✓ Including `tank` as a blocking factor removes the **between-tank variation in background growth conditions** from the residual error term. This variation (due to differences in temperature, flow rate, light, or other unmeasured tank-level factors) would otherwise inflate the residual mean square and mask treatment effects.
- ✓ Statistically: the block effect partitions tank-to-tank variation from the residual, making the denominator of the *F*-ratio for treatment smaller — the signal-to-noise ratio for detecting treatment effects improves.

d.

- ✓ Without the blocking factor, the between-tank variance would be lumped into the residual error (pooled with the within-treatment variation). If tank explains substantial variance (as suggested by the consistent within-tank differences in the data preview), the residual mean square would **increase substantially**, inflating the *F*-ratio denominator.
- ✓ The *F*-statistic for treatment would be **reduced** (treatment mean square / larger residual), reducing power to detect the nutrient effect. The RCBD ANOVA (with blocking) will produce a smaller residual and a larger, more powerful *F*-test for the treatment of interest.

Part C: Statistical Output Interpretation (37 marks)

Question 13 — Poisson GLM Output (/12)

An entomologist models the number of moth species observed per forest plot (n = 80 plots) as a function of tree species diversity (Simpson's D, continuous) and forest age (years, continuous). The Poisson GLM output is:

```
Call:
glm(formula = moth_species ~ simpson_D + forest_age,
    family = poisson, data = forest_moths)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.8420     0.2130    8.65  < 0.001 ***
simpson_D      1.2340     0.2810    4.39  < 0.001 ***
forest_age     0.0083     0.0021    3.95  < 0.001 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 187.4  on 79  degrees of freedom
Residual deviance:  98.3  on 77  degrees of freedom
AIC: 412.7
```

a. Why is a Poisson GLM more appropriate than a standard linear model for this response variable? **(/ 2)**

b. Interpret the coefficient for `simpson_D` (1.2340) on the **response (count) scale**. **(/ 3)**

c. The residual deviance (98.3) on 77 df gives a dispersion ratio of approximately 1.28. What does this suggest, and would you be concerned? **(/ 3)**

d. The null deviance is 187.4 and the residual deviance is 98.3. Calculate the **percentage of deviance explained** by the model and interpret this value. **(/ 4)**

Tip

Model Answer — Question 13

a.

- ✓ The response variable (moth species richness) consists of **non-negative integer counts**. A standard linear model (Gaussian errors) can predict negative values, assumes normally distributed residuals with constant variance, and is not designed for count data.
- ✓ A Poisson GLM is appropriate because it models the **expected count** via a log link (ensuring positivity), uses the Poisson distribution (where variance scales with the mean), and is designed for count responses.

b.

- ✓ On the **log scale**: for a one-unit increase in Simpson's D (tree diversity), log(expected moth species count) increases by 1.234.
- ✓ On the **response (count) scale**: the expected moth species count is multiplied by $e^{1.234} \approx$ **3.44** for each one-unit increase in Simpson's D — a 244% increase in expected moth species count.
- ✓ Biologically: plots with higher tree diversity support approximately 3.4 times as many moth species as plots with one unit lower tree diversity (holding forest age constant). This large multiplicative effect is consistent with the resource diversity hypothesis: diverse tree communities provide more host plant species and microhabitat types, supporting a richer moth community.

c.

- ✓ A dispersion ratio (residual deviance / residual df) of 98.3 / 77 ≈ **1.28** is somewhat above 1 but is **mild** and not severely alarming. In a well-fitted Poisson model, this ratio should be approximately 1.0; values modestly above 1 may reflect slight overdispersion.
- ✓ A dispersion ratio of 1.28 is within a range that many practitioners would accept without immediate concern, but it suggests the model could be refined — a **quasi-Poisson** model (adjusting standard errors for slight overdispersion) or a **negative binomial GLM** (if overdispersion is attributed to unmeasured heterogeneity among plots) would be more conservative alternatives.

d.

- ✓ Percentage of deviance explained = (Null deviance − Residual deviance) / Null deviance × 100 = (187.4 − 98.3) / 187.4 × 100 = 89.1 / 187.4 × 100 ≈ **47.5%**.
- ✓ This is analogous to R^2 in ordinary regression: the model with tree diversity and forest age explains approximately **47.5% of the null deviance** (the total variation in moth species richness among plots).
- ✓ This represents a moderate-to-strong explanatory performance — roughly half the variation in moth diversity is accounted for by these two predictors, with the remaining ~~~52.5% attributable to unmeasured environmental factors, spatial heterogeneity, or stochas-~~ tic processes.

Question 14 — Repeated Measures ANOVA Output (/12)

A plant ecologist measures leaf chlorophyll content (SPAD units) in 20 individual *Phragmites australis* plants at four phenological stages: Early growth (E), Peak growth (P), Senescence (S), and Dormancy (D). The same plants are measured at each stage. The repeated measures ANOVA table is:

```
                   Df  Sum Sq  Mean Sq  F value   Pr(>F)
phenological_stage  3  2418.7   806.2   47.83   < 0.001 ***
Residuals          57   960.8    16.9

Mauchly's test of sphericity: W = 0.812, p-value = 0.063
```

a. State the null hypothesis evaluated by the F-test. **(/ 2)**
b. Interpret the F-value (47.83) and its associated p-value. **(/ 3)**
c. The Mauchly's test returns $p = 0.063$. What does this test evaluate, and what action, if any, should the researcher take? **(/ 3)**
d. If the analysis is significant, what would be the appropriate **post-hoc approach**, and how many pairwise comparisons are involved? **(/ 2)**
e. What does the residual mean square (16.9) represent in this repeated measures design? **(/ 2)**

Tip

Model Answer — Question 14

a.

- ✓ H_0: Mean leaf chlorophyll content (SPAD) does not differ among the four phenological stages: $\mu_E = \mu_P = \mu_S = \mu_D$. The repeated time-course of chlorophyll shows no systematic change across the growth cycle.

b.

- ✓ $F(3, 57) = 47.83$ means that the between-stage variance is 47.83 times greater than the within-subject (residual) variance — the phenological stage differences far exceed what would be expected from random within-plant variation alone.
- ✓ $p < 0.001 \ll \alpha = 0.05$: we **strongly reject** H_0. There is overwhelming statistical evidence that mean chlorophyll content changes significantly across phenological stages in *P. australis*.
- ✓ However, the omnibus *F*-test only establishes that *some* difference exists; post-hoc testing is required to identify which specific stage pairs differ.

c.

- ✓ **Mauchly's test of sphericity** evaluates whether the variances of the differences between all pairs of within-subjects conditions are equal — the sphericity assumption required for the *F*-test in repeated measures ANOVA to be valid.
- ✓ Here, $p = 0.063 > 0.05$: we fail to reject the sphericity null hypothesis at $\alpha = 0.05$ — **marginal evidence** of a violation. Many researchers would interpret this conservatively and apply a **Greenhouse-Geisser correction** to the degrees of freedom to guard against inflation of the Type I error rate, given that the test is borderline.
- ✓ Practical action: report both the uncorrected result and the GG-corrected *F*-test to demonstrate robustness. The correction reduces the effective df (making the test slightly more conservative), which is prudent when sphericity is marginal.

d.

- ✓ Apply **pairwise paired *t*-tests with Bonferroni (or Holm) correction** to all pairwise combinations of phenological stages.
- ✓ Number of pairwise comparisons: $k(k-1)/2 = 4 \times 3 / 2 =$ **6 comparisons** (E vs. P, E vs. S, E vs. D, P vs. S, P vs. D, S vs. D). The Bonferroni-adjusted threshold would be $\alpha / 6 \approx 0.0083$.

e.

- ✓ In a repeated measures design, the residual mean square (16.9) represents the **within-subject variability** — the average squared deviation of individual plants' measurements from their own mean trajectory across stages, after accounting for the stage effect.
- ✓ It reflects how consistently each plant changes across stages: plants that show highly consistent chlorophyll dynamics contribute little to the residual; plants with irregular responses inflate it. This within-subject error is substantially smaller than the between-subject variance (individual differences in mean chlorophyll), which is why the repeated measures design is powerful.

Question 15 — Multiple Regression Output: Kelp Frond Length (/13)

A kelp ecologist models the frond length (cm) of *Ecklonia maxima* at 75 sampling locations as a function of three continuous environmental predictors: daily irradiance (mol photons m^{-2} day^{-1}), water temperature (°C), and nitrate concentration (µmol L^{-1}). The `lm()` output is:

```
Call:
lm(formula = frond_length_cm ~ irradiance + temperature + nitrate,
   data = kelp)

Residuals:
    Min      1Q  Median      3Q     Max
 -28.43  -7.91    0.34   8.12   31.17

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -45.240      8.120   -5.57  < 0.001 ***
irradiance      3.840      0.530    7.25  < 0.001 ***
temperature     2.120      0.680    3.12   0.0026 **
nitrate         1.470      0.410    3.59   0.0006 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.34 on 71 degrees of freedom
Multiple R-squared:  0.7401,  Adjusted R-squared:  0.7291
F-statistic: 67.42 on 3 and 71 DF,  p-value: < 2.2e-16
```

a. Write the fitted regression equation and comment on whether the intercept is biologically interpretable. **(/ 2)**

b. Interpret the coefficient for `irradiance` (3.840) as a **partial regression coefficient**. What does "partial" mean in this context? **(/ 3)**

c. A colleague argues that `temperature` should be removed from the model because its p-value (0.0026) is larger than those of the other two predictors. Evaluate this argument. **(/ 3)**

d. What does adjusted $R^2 = 0.7291$ indicate, and why is it lower than $R^2 = 0.7401$? **(/ 3)**

e. A new site has irradiance = 8.5 mol photons m^{-2} day^{-1}, temperature = 16.2°C, and nitrate = 4.8 µmol L^{-1}. Calculate the predicted frond length. **(/ 2)**

Tip

Model Answer — Question 15

a.

- ✓ $\widehat{frond\_length} = -45.240 + 3.840 \times irradiance + 2.120 \times temperature + 1.470 \times nitrate$
- ✓ The intercept (−45.24 cm) is the predicted frond length when all three predictors simultaneously equal zero — a combination that never occurs in nature (zero irradiance, zero temperature, zero nitrate). It is a **mathematical anchor** for the regression plane, not a biologically meaningful quantity; extrapolation to zero values is outside the observed data range.

b.

- ✓ The coefficient 3.840 is a **partial regression coefficient**: it estimates the change in expected frond length associated with a one-unit increase in irradiance (1 mol photon m^{-2} day^{-1}), **holding temperature and nitrate constant**. For each additional mol photon m^{-2} day^{-1} of irradiance, kelp fronds are predicted to be 3.84 cm longer, all else being equal.
- ✓ "Partial" means this estimate isolates the unique contribution of irradiance to frond length *after* accounting for the linear effects of temperature and nitrate — it is not the simple (marginal) effect of irradiance in isolation, which would also absorb any variation shared with the other two predictors.
- ✓ The strong positive relationship is biologically consistent: greater light availability drives photosynthesis and carbon fixation, fuelling frond elongation growth.

c.

- ✓ The colleague's argument is **incorrect**. All three predictors are statistically significant at $\alpha = 0.05$ (temperature: $p = 0.0026$). A lower p-value does not mean one predictor is "better" or that others should be dropped — it reflects a combination of effect size and estimation precision.
- ✓ Model selection should be guided by **AIC**, **adjusted R^2**, or biological reasoning — not by p-value ranking among retained predictors. Removing a significant predictor (temperature, $p = 0.0026 \ll 0.05$) without justification would discard a real effect and bias the remaining coefficients if temperature is correlated with irradiance or nitrate (omitted variable bias).
- ✓ Furthermore, each predictor's p-value already accounts for all other predictors in the model. Temperature remains significant conditional on irradiance and nitrate — it has an independent contribution to frond length that would be conflated with the other predictors if removed.

d.

- ✓ Adjusted $R^2 = 0.7291$ means that approximately **72.9% of the variation in kelp frond length** is explained by the three-predictor model, after penalising for model complexity (number of predictors relative to sample size).
- ✓ Adjusted R^2 is lower than R^2 (0.7401) because it applies a **penalty for each additional parameter** estimated: $R^2_{adj} = 1 - (1 - R^2) \times (n - 1) / (n - k - 1)$, where $n = 75$ and $k = 3$. Unlike R^2, adjusted R^2 does not automatically increase when a predictor is added — it decreases if the predictor adds less explanatory power than expected by chance. The small gap (0.0110) here confirms that all three predictors are genuinely contributing.

e.

- ✓ $\widehat{frond\_length} = -45.240 + 3.840 \times 8.5 + 2.120 \times 16.2 + 1.470 \times 14.8$

End of Version 6

Bibliography