# BCB744 Biostatistics — Theory Test (Version 10)

## 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 — Confounding Variables and Study Design (/6)

a. Define a **confounding variable** and explain why it is a threat to valid causal inference. Illustrate your answer with a biological example in which a plausible confound explains an observed association between two variables. **(/ 3)**

b. Explain how **random assignment** of subjects to treatment groups controls for confounding variables, and why this control is generally stronger than **statistical adjustment** (e.g., including the confound as a covariate). **(/ 3)**

Tip

**Model Answer — Question 1**

*a.*

- ✓ A **confounding variable** is a third variable that is associated with both the explanatory variable and the response variable. Because the confounder changes along with the explanatory variable, its effect is entangled with the effect of interest, making it impossible to determine whether the observed association is causal or merely spurious.
- ✓ Example: A survey finds that rocky-shore sites with more macroalgae also have more limpets. Confound: wave exposure. Sheltered sites tend to support dense algal mats *and* large limpet populations, so the macroalgae–limpet association may reflect shared exposure conditions rather than a direct relationship.
- ✓ Without controlling for wave exposure, any conclusion that macroalgae drive limpet abundance would be unwarranted.

*b.*

- ✓ **Random assignment** distributes both measured *and* unmeasured confounders approximately equally across treatment groups in expectation. Because each subject has an equal probability of receiving any treatment, the groups should be comparable on all background variables — including ones the researcher never measured. This removes systematic bias.
- ✓ **Statistical adjustment** (covariate control) only removes the influence of confounders that the researcher identified and measured correctly. It cannot control for unknown confounders, and it requires correct model specification (linearity, no interaction with treatment). Random assignment is therefore a stronger guarantee of causal inference because it does not depend on knowing what to measure.

## Question 2 — Descriptive Statistics: Skewness and Outliers (/7)

a. Distinguish between **positively skewed** and **negatively skewed** distributions. For each, state whether the mean lies above or below the median and explain why. **(/ 3)**

b. A marine ecologist measures body mass (g) of 40 individual anchovies and obtains: min = 12.1, Q1 = 18.4, median = 21.7, Q3 = 25.3, max = 68.9. Using the **IQR rule**, determine whether the maximum value is a potential outlier. Show your working. **(/ 2)**

c. A colleague proposes reporting the **mean** body mass for the anchovy dataset. Given your findings in (b), argue whether the mean or the median is a more appropriate measure of central tendency here, and explain the reasoning. **(/ 2)**

Tip

**Model Answer — Question 2**

*a.*

- ✓ **Positively skewed** (right-skewed): the tail extends toward high values. A small number of very large values pull the mean upward but have less effect on the median (which depends only on rank). Therefore **mean > median**.
- ✓ **Negatively skewed** (left-skewed): the tail extends toward low values. A small number of very low values pull the mean downward. Therefore **mean < median**.
- ✓ The median is resistant to extreme values because it is determined by the middle rank; the mean is sensitive to all values and is dragged toward the long tail.

*b.*

- ✓ IQR = Q3 − Q1 = 25.3 − 18.4 = 6.9.
- ✓ Upper fence = Q3 + 1.5 × IQR = 25.3 + 1.5 × 6.9 = 25.3 + 10.35 = **35.65 g**.
- ✓ The maximum value of 68.9 g exceeds 35.65 g, so it **is** a potential outlier by the IQR rule.

*c.*

- ✓ Because a potential outlier (68.9 g) is present and the data are likely right-skewed (max is far above Q3 but Q1 is relatively close to the median), the mean will be inflated by the extreme value. The **median** (21.7 g) is a more robust and representative measure of central tendency for this dataset.
- ✓ Alternatively: if the outlier is verified as a genuine biological measurement rather than an error, one could retain the mean but report it alongside the median and note that the distribution is right-skewed.

## Question 3 — Data Visualisation (/7)

a. A researcher wants to compare the distributions of seagrass shoot density (shoots $m^{-2}$) among **five** sites, each with $n = 25$ replicate quadrats. List **two** appropriate visualisation types, and explain what each reveals that a bar chart of means ± SE does *not*. **(/ 3)**

b. Explain why a **log-scale** y-axis is sometimes preferable to a linear scale when plotting biological abundance data across multiple orders of magnitude. What is one potential disadvantage of using a log scale? **(/ 2)**

c. A scatter plot of algal biomass (g $m^{-2}$) against light availability (µmol photons $m^{-2}$ $s^{-1}$) shows clear curvature. Describe the transformation(s) you would apply to the data (or the model) to address this and explain the reasoning. **(/ 2)**

Tip

**Model Answer — Question 3**

*a.*

- ✓ **Box plot**: shows median, IQR, whiskers, and individual outliers. Reveals the spread, symmetry/skewness, and extreme values for each site — information hidden when only mean ± SE is shown.
- ✓ **Violin plot** (or strip/dot plot): shows the full empirical distribution (or individual data points). Reveals multimodality, gaps, and the actual distribution shape that box plots summarise but box plots themselves can obscure.
- ✓ Bar charts of means ± SE conceal the sample distribution, potential outliers, and whether the data are symmetric or skewed.

*b.*

- ✓ When data span multiple orders of magnitude (e.g., 1 to 10,000 individuals), a linear scale compresses the low end into an indistinguishable band. A log scale spreads all orders of magnitude equally, making patterns at low and high values simultaneously visible.
- ✓ Disadvantage: a log scale cannot display zero values (log(0) is undefined), which is common in biological count data. It also makes multiplicative (percentage) differences appear as additive differences, which can be misread by audiences unfamiliar with log scales.

*c.*

- ✓ The curvature suggests a **nonlinear relationship**. Appropriate approaches: (i) log-transform or square-root-transform the predictor (light availability) to linearise the relationship; (ii) add a **polynomial term** (quadratic) for light in the regression model; or (iii) log-transform the response if the relationship follows a power law.
- ✓ The reasoning: OLS regression assumes a linear relationship between predictor and response. Curvature violates this assumption, inflating residuals in a structured (non-random) pattern. Linearising the relationship (via transformation or polynomial extension) satisfies the linearity assumption and improves model fit.

## Question 4 — The Normal Distribution and *z*-Scores (/6)

a. State the **empirical rule** (68–95–99.7 rule) for a normal distribution. A seabird lays eggs with mean mass $\mu = 85$ g and $\sigma = 6$ g. What percentage of eggs would you expect to have a mass between 73 g and 97 g? Show reasoning. **(/ 3)**

b. A single egg is found with a mass of 100 g. Calculate its **z-score** and interpret the result in plain language (i.e., what does this z-score tell you about the egg relative to the population?). **(/ 3)**

Tip

**Model Answer — Question 4**

*a.*

- ✓ The empirical rule: approximately 68% of observations fall within ±1σ of μ; approximately 95% within ±2σ; approximately 99.7% within ±3σ.
- ✓ 73 g = 85 − 2 × 6 = μ − 2σ; 97 g = 85 + 2 × 6 = μ + 2σ. The interval [73, 97] spans exactly ±2σ.
- ✓ Therefore approximately **95%** of eggs are expected to have masses in this range.

*b.*

- ✓ $z = (100 - 85) / 6 = 15 / 6 =$ **+2.50**.
- ✓ A $z$-score of +2.50 means the egg is **2.5 standard deviations above the population mean**. This is a relatively rare observation: only about 1.24% of eggs (from standard normal tables, $P(Z > 2.50) \approx 0.0062$, so ~0.62% in one tail, ~1.24% more extreme in absolute value) would be expected to be this far or farther from the mean.
- ✓ In biological terms: this egg is unusually large relative to the population.

## Question 5 — One-Tailed vs. Two-Tailed Tests and Effect Size (/7)

a. Distinguish between a **one-tailed** and a **two-tailed** hypothesis test. Under what conditions is a one-tailed test scientifically justified, and what is the main risk of using it without justification? **(/ 3)**

b. Define **Cohen's *d*** as a measure of effect size and explain what it quantifies that a $p$-value does not. A study comparing foraging success of male vs. female albatrosses finds $t(58) = 2.04$, $p = 0.046$, $d = 0.52$. Interpret both statistics and explain whether the effect is scientifically meaningful. **(/ 4)**

Tip

**Model Answer — Question 5**

*a.*

- ✓ A **two-tailed** test evaluates whether the parameter differs from the null in *either* direction ($H_A$: $\mu_1 \neq \mu_2$); the significance level $\alpha$ is split equally between both tails. A **one-tailed** test evaluates a directional alternative ($H_A$: $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$); the full $\alpha$ is placed in one tail, giving greater power to detect effects in the predicted direction.
- ✓ A one-tailed test is scientifically justified when prior theory or knowledge gives a *strong, directional a priori* expectation — for example, testing whether a toxicant *reduces* (not merely changes) survival. The direction must be specified before data collection.
- ✓ Risk: using a one-tailed test without prior justification is a form of *p*-hacking — it halves the effective threshold needed to reject $H_0$, inflating the Type I error rate beyond the nominal $\alpha$.

*b.*

- ✓ **Cohen's *d*** = $(\mu_1 - \mu_2) / SD_{pooled}$. It expresses the difference between group means in units of the pooled standard deviation, giving a **standardised, scale-free measure of the magnitude of the difference**. By convention: $d \approx 0.2$ is small, $d \approx 0.5$ is medium, $d \approx 0.8$ is large.
- ✓ The *p*-value ($p = 0.046$) tells us that the observed difference is statistically significant at $\alpha = 0.05$ — i.e., unlikely to have arisen by chance if $H_0$ is true. It does **not** indicate the size or practical importance of the difference.
- ✓ $d = 0.52$ is a **medium effect size**: the sex difference in foraging success is moderate in standardised terms. Given a statistically significant result with a medium effect, the difference is both unlikely due to chance and of biological relevance. The researcher should consider the biological context (e.g., does a 0.52 SD difference in foraging success affect fitness?) to assess practical importance.

## Question 6 — Assumptions and Data Transformations (/7)

a. List the **four main assumptions** of the independent-samples *t*-test. For each, name one diagnostic method (graphical or statistical) used to assess it. **(/ 4)**

b. A researcher wants to compare bacterial colony counts between two treatment groups. The data are heavily right-skewed with variances that differ between groups by a factor of 8. Which **single transformation** would most likely address both problems simultaneously, and why? **(/ 2)**

c. After applying the transformation from (b), the researcher still detects a significant difference between the groups. When reporting results, should the researcher report the back-transformed mean or the mean on the transformed scale? Justify your answer. **(/ 1)**

Tip

**Model Answer — Question 6**

*a.*

- ✓ **Independence of observations**: no built-in diagnostic, but checked via study design (random sampling, no repeated measurements on the same subject).
- ✓ **Normality of residuals** (or of each group): assessed with a **Q-Q plot** or a Shapiro-Wilk test on each group.
- ✓ **Homogeneity of variance** (equal variances): assessed with a **Levene's test** or by comparing the ratio of the larger to smaller sample variance (ratio > ~4 is concerning).
- ✓ **Continuous (at least interval-level) measurement scale**: confirmed by understanding the nature of the variable — assessed by knowing the data type.

*b.*

- ✓ A **log-transformation** ($\log_{10}$ or natural log) is most appropriate. Right-skewed count data are multiplicatively structured: large values differ from small values by a ratio rather than a fixed difference. Taking the log converts multiplicative relationships to additive ones, simultaneously compressing the right tail (reducing skewness) and stabilising variance (because the standard deviation of counts is often proportional to the mean — the log transformation corrects this heteroscedasticity).

*c.*

- ✓ The researcher should report the **back-transformed mean** (i.e., 10^mean_log or e^mean_log) when communicating results to a biological audience, as it represents the **geometric mean** on the original scale and is more interpretable. However, it is good practice to note that statistical tests were performed on log-transformed data and to report the result on the transformed scale in the methods/results.

## Question 7 — Independent-Samples *t*-Test and Welch's Correction (/6)

a. Explain the difference between **Student's independent-samples *t*-test** and **Welch's *t*-test**. When is Welch's test preferred, and what is the cost of using it? **(/ 3)**

b. A study compares the leaf area index (LAI) of two fynbos shrub species measured at 15 sites each. The standard deviations are 0.42 (species A) and 1.31 (species B). State whether Student's or Welch's test is more appropriate and explain your reasoning. Would a non-parametric alternative be warranted? Under what condition? **(/ 3)**

Tip

**Model Answer — Question 7**

*a.*

- ✓ **Student's *t*-test** assumes equal population variances (homoscedasticity) and pools the two sample variances to estimate a single common variance. Welch's *t*-test does **not** assume equal variances; it uses separate variance estimates for each group and adjusts the degrees of freedom downward using the Welch-Satterthwaite equation.
- ✓ Welch's test is preferred when group variances are noticeably unequal (Levene's test significant, or variance ratio > ~4). It is more robust to heteroscedasticity while maintaining good Type I error control.
- ✓ The cost: Welch's test has **lower degrees of freedom** (by the Satterthwaite correction), giving slightly less power than Student's *t*-test when variances truly are equal. In practice this loss is small, and many statisticians recommend always using Welch's *t*-test as the default.

*b.*

- ✓ The variance ratio is $(1.31)^2$ / $(0.42)^2$ ≈ 1.716 / 0.176 ≈ **9.75**, which substantially exceeds the guideline of ~4. The homoscedasticity assumption of Student's *t*-test is likely violated. **Welch's *t*-test** is more appropriate.
- ✓ A non-parametric alternative (Wilcoxon rank-sum test) would be warranted if, in addition to unequal variances, the data within one or both groups are severely non-normal — which could be checked with Q-Q plots or a Shapiro-Wilk test. With $n$ = 15 per group, the CLT provides less protection, so normality of the data matters more.

## Question 8 — Interaction Effects in Two-Way ANOVA (/8)

a. Define a **statistical interaction** in the context of a two-way factorial ANOVA. Explain what it means biologically when an interaction is significant. **(/ 3)**

b. A researcher tests the effects of **temperature** (15°C, 20°C, 25°C) and **salinity** (30 ppt, 35 ppt) on the photosynthetic rate (μmol $O_2$ mg $chl^{-1}$ $h^{-1}$) of a marine microalga in a fully crossed 3 × 2 factorial design. The ANOVA table reveals a significant interaction term ($F(2, 54) = 8.73$, $p < 0.001$).

   Sketch (in text or a simple table) an **interaction plot** that would reflect this result, with temperature on the x-axis and separate lines for each salinity level. Explain what feature of the plot indicates an interaction. **(/ 3)**

c. Given a significant interaction, why is it **misleading** to interpret the main effects of temperature and salinity in isolation? What should the researcher report instead? **(/ 2)**

Tip

**Model Answer — Question 8**

*a.*

- ✓ A **statistical interaction** occurs when the effect of one factor on the response variable **depends on the level of another factor**. In other words, the factors do not act independently — their joint effect is not simply the sum of their individual effects.
- ✓ Biologically: the temperature × salinity interaction means that the effect of temperature on photosynthetic rate is **different** at 30 ppt compared to 35 ppt. For example, photosynthesis might increase sharply with temperature at low salinity (indicating thermal stimulation under osmotic stress), but remain flat or decline at high salinity — the two stressors interact rather than act independently.

*b.*

- ✓ In an interaction plot, the lines for the two salinity levels would **not be parallel** — they converge, diverge, or cross. For example:

| Temperature | 30 ppt | 35 ppt |
|---|---|---|
| 15°C | 4.2 | 3.8 |
| 20°C | 6.5 | 5.1 |
| 25°C | 7.8 | 4.4 |

  The line for 30 ppt rises steeply from 15°C to 25°C, while the line for 35 ppt rises only slightly and then levels off. The **non-parallel (diverging) lines** indicate an interaction.

- ✓ If the lines were parallel (same slope for both salinity levels), there would be no interaction: temperature would have the same effect regardless of salinity.

*c.*

- ✓ When an interaction is significant, the main effect of each factor is an average over levels of the other factor — an average that is **not meaningful** if the factor's effect changes direction or magnitude depending on the other factor's level. Reporting only main effects implies a uniform effect of temperature across all salinities (or vice versa), which is false.
- ✓ The researcher should report the **simple effects** (i.e., the effect of temperature at each salinity level separately, and the effect of salinity at each temperature level), and display the interaction plot to visualise the conditional nature of each factor's effect.

## Question 9 — Collinearity and VIF in Multiple Regression (/7)

a. Explain what **collinearity** (multicollinearity) means in the context of multiple regression and describe **two consequences** it has for the estimated regression coefficients. **(/ 3)**

b. Define the **Variance Inflation Factor (VIF)** and explain what it measures. What rule-of-thumb threshold(s) are commonly used to flag problematic collinearity? **(/ 2)**

c. A multiple regression model of fish biomass (g) includes the predictors: water temperature (°C), salinity (ppt), and depth (m). The VIFs are: temperature = 1.28, salinity = 8.41, depth = 7.93. What do these values suggest, and what actions might the researcher take to address the issue? **(/ 2)**

Tip

**Model Answer — Question 9**

*a.*

- ✓ **Collinearity** occurs when two or more predictor variables in a multiple regression model are highly correlated with each other. The model cannot reliably separate the individual contributions of each predictor to the response.
- ✓ Consequence 1: **Inflated standard errors** — the sampling uncertainty of each coefficient increases markedly, making it harder to detect genuine effects (reduced power, wider confidence intervals).
- ✓ Consequence 2: **Unstable coefficient estimates** — small changes in the data (adding or removing a few observations) can produce large swings in the estimated coefficients, making them unreliable and difficult to interpret.

*b.*

- ✓ VIF for predictor $j$ = 1 / (1 − $R^2$_j), where $R^2$_j is the $R^2$ from regressing predictor $j$ on all other predictors. It quantifies **how much the variance of the estimated coefficient for predictor *j* is inflated due to its correlation with the other predictors**. A VIF of 1 indicates no collinearity; higher values indicate increasing collinearity.
- ✓ Common thresholds: VIF > **5** warrants concern; VIF > **10** is generally considered severe collinearity requiring action.

*c.*

- ✓ Temperature's VIF (1.28) is unproblematic. Salinity (8.41) and depth (7.93) both exceed 5 and are close to 10 — suggesting that **salinity and depth are highly correlated** with each other (or with temperature together), inflating the uncertainty of their coefficients.
- ✓ Possible actions: (i) **remove one** of the correlated predictors (retaining the one with stronger biological rationale); (ii) **combine** salinity and depth into a single composite variable (e.g., via PCA) if both convey related information; (iii) collect additional data that better separates salinity and depth variation; (iv) centre and standardise predictors and re-check VIFs. The researcher should also inspect pairwise correlations among predictors to identify which specific pair is responsible.

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

## Question 10 — Paired *t*-Test: Thermal Stress in Coral Recruits (/12)

A marine physiologist investigates whether a two-week heat stress event (temperature elevated by +2°C above ambient) affects the **chlorophyll *a* concentration** (μg $cm^{-2}$) in juvenile coral recruits as a proxy for symbiont (Symbiodiniaceae) density. Twenty coral fragments from the same parent colony are used; each fragment is split in two, with one half assigned to the control aquarium (ambient temperature) and the other half to the heated aquarium. Chlorophyll *a* is measured in each half-fragment after two weeks.

a. Identify the **experimental unit** in this study and explain why a **paired *t*-test** is appropriate rather than an independent-samples *t*-test. **(/ 3)**
b. State the null and alternative hypotheses for this study using correct statistical notation. **(/ 2)**
c. The researcher checks assumptions before proceeding. Describe the **two key assumptions** specific to the paired *t*-test (beyond those shared with all parametric tests) and explain how each would be assessed. **(/ 3)**
d. The results show: mean difference = −1.84 μg $cm^{-2}$, SD of differences = 2.31 μg $cm^{-2}$, $n$ = 20 pairs. Calculate the **paired *t*-statistic** and the degrees of freedom. Interpret the result in plain language (you do not need to look up a critical value — state what additional information would be needed to make a decision). **(/ 4)**

Tip

**Model Answer — Question 10**

*a.*

- ✓ The **experimental unit** is the **coral fragment** (or more precisely, the pair of half-fragments). Each fragment contributes one control measurement and one heated measurement.
- ✓ The paired design is appropriate because the two measurements within each pair are **not independent**: both come from the same parent fragment and therefore share genetic and physiological background. Using an independent-samples *t*-test would ignore this within-pair correlation and waste statistical information. The paired *t*-test accounts for within-pair similarity by analysing the *differences*, removing between-fragment variability from the error term and increasing power.

*b.*

- ✓ $H_0$: μ_d = 0 (the mean difference in chlorophyll *a* between control and heated half-fragments is zero; heat stress has no effect).
- ✓ $H_A$: μ_d ≠ 0 (the mean difference is not zero; heat stress changes chlorophyll *a* concentration). *(A one-tailed $H_A$: μ_d < 0 is also acceptable if justified by prior evidence that heat bleaches corals.)*

*c.*

- ✓ **The differences are approximately normally distributed**: since the *t*-test is applied to the within-pair differences, normality of the differences (not the raw data) is required. Assessed with a Q-Q plot of the differences, or a Shapiro-Wilk test on the $n = 20$ difference values.
- ✓ **The pairs are independent of each other**: the difference observed for one fragment must not influence the difference for another. Assessed by study design — fragments from the same colony are acceptable as long as they are not in the same aquarium or otherwise cross-contaminated.

*d.*

- ✓ $t = \bar{d} / (SD\_d / \sqrt{n}) = -1.84 / (2.31 / \sqrt{20}) = -1.84 / (2.31 / 4.472) = -1.84 / 0.5166 =$ **−3.56**.
- ✓ Degrees of freedom = $n - 1 = 20 - 1 =$ **19**.
- ✓ Interpretation: the paired *t*-statistic is −3.56 on 19 df. This means the mean chlorophyll *a* was 1.84 μg cm$^{-2}$ lower in heated fragments, and the difference is 3.56 standard error units below zero. To make a formal decision, we would need to compare the *t*-statistic to the critical value at $t_{0.025, 19} \approx 2.093$ (two-tailed, $\alpha = 0.05$). Since $|-3.56| > 2.093$, the result would be significant: we would reject $H_0$ and conclude that heat stress significantly reduced chlorophyll *a* concentration.

## Question 11 — One-Way ANOVA and Kruskal-Wallis: Polychaete Assemblages (/13)

A benthic ecologist samples polychaete worm species richness (number of species per core) across **four** subtidal sediment habitats: mud, fine sand, coarse sand, and gravel. Fifteen cores are taken in each habitat ($n$ = 15 per group, $N$ = 60 total). The data are count-like, with several zeros in the mud habitat; Q-Q plots show marked right-skew and the Shapiro-Wilk test is significant for the mud and fine sand groups ($p < 0.05$).

a. Given the distributional findings, state whether a **one-way ANOVA** or a **Kruskal-Wallis test** is more appropriate, and justify your choice. **(/ 3)**
b. State the null and alternative hypotheses for the chosen test in the context of this study. **(/ 2)**
c. The Kruskal-Wallis test returns $H = 18.42$, df = 3, $p = 0.00036$. Interpret this result. What does the significant $p$-value tell you, and what does it **not** tell you? **(/ 3)**
d. A post-hoc analysis (Dunn test with Bonferroni correction) is applied. With four groups, how many pairwise comparisons are performed, and what is the Bonferroni-corrected significance threshold if the family-wise error rate is to be kept at $\alpha = 0.05$? Explain why this correction is necessary. **(/ 3)**
e. The ecologist reports that gravel habitat has significantly higher polychaete richness than mud ($p_{\text{adjusted}} = 0.0021$). Write a **one-sentence conclusion** suitable for the results section of a scientific paper. **(/ 2)**

Tip

**Model Answer — Question 11**

*a.*

- ✓ The **Kruskal-Wallis test** is more appropriate. One-way ANOVA assumes approximately normal distributions within groups and homogeneous variances; here, the Shapiro-Wilk test is significant for two groups and the data are right-skewed with zeros. With $n = 15$ per group, the CLT provides limited protection against non-normality. The Kruskal-Wallis test, as a non-parametric alternative that operates on ranks, is robust to non-normality and unequal variances.

*b.*

- ✓ $H_0$: The distributions of polychaete species richness are identical across all four sediment habitats (all groups have the same population median/distribution).
- ✓ $H_A$: At least one sediment habitat has a different distribution of polychaete species richness compared to the others.

*c.*

- ✓ The Kruskal-Wallis statistic $H = 18.42$ on 3 df gives $p = 0.00036 < 0.05$. We **reject $H_0$** and conclude that polychaete species richness differs significantly among at least some of the four sediment habitats.
- ✓ The significant $p$-value does **not** tell us *which* specific habitats differ from each other — only that the groups are not all the same. Post-hoc pairwise comparisons are needed to identify which pairs differ.

*d.*

- ✓ With $k = 4$ groups, the number of pairwise comparisons = $k(k-1)/2 = 4 \times 3 / 2 = \mathbf{6}$.
- ✓ Bonferroni-corrected threshold = $\alpha / m = 0.05 / 6 \approx \mathbf{0.0083}$.
- ✓ The correction is necessary because conducting multiple tests inflates the family-wise Type I error rate. If six independent tests are each conducted at $\alpha = 0.05$, the probability of at least one false positive is $1 - (0.95)^6 \approx 0.26$. The Bonferroni correction keeps the overall false-positive rate at or below 5%.

*e.*

- ✓ "Polychaete species richness was significantly higher in gravel habitat than in mud habitat (Kruskal-Wallis test with Bonferroni-corrected post-hoc Dunn test, $p_{\text{adjusted}} = 0.0021$)."

## Question 12 — Multiple Regression with Interaction: Fynbos Shrub Growth (/12)

A plant ecologist studies height growth (cm year$^{-1}$) of the fynbos shrub *Protea repens* across 80 sites in the Western Cape. She measures two continuous predictors: **mean annual rainfall** (mm

year$^{-1}$; centred to $x_1$) and **soil phosphorus** (mg kg$^{-1}$; centred to $x_2$). She fits a multiple regression model **with** an interaction term:

$$\widehat{\text{growth}} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2$$

a. Explain why the ecologist **centred** the continuous predictors before fitting the interaction model, and what would happen to the interpretation of $b_1$ and $b_2$ if centring were not applied. **(/ 3)**

b. State the null and alternative hypotheses for the **interaction term** ($b_3$) and explain in biological terms what a significant interaction would mean. **(/ 3)**

c. The fitted model yields: $b_0 = 4.82$, $b_1 = 0.031$, $b_2 = -0.14$, $b_3 = 0.0018$. Interpret each coefficient in biological terms, including the intercept. **(/ 4)**

d. A colleague argues that the interaction term should be dropped because it adds complexity. What statistical criterion would you use to decide whether the interaction term is worth retaining, and what would you look at? **(/ 2)**

Tip

**Model Answer — Question 12**

*a.*

- ✓ **Centring** subtracts the mean of each predictor from every observation ($x_1$ = rainfall − mean rainfall; $x_2$ = phosphorus − mean phosphorus). When an interaction term ($x_1x_2$) is included, the product of two uncentred variables is often highly correlated with the individual predictors (collinearity), inflating VIFs and making it difficult to estimate the main-effect coefficients reliably.
- ✓ After centring, $b_1$ and $b_2$ represent the **marginal effects** of rainfall and phosphorus *at the mean value of the other predictor* (i.e., when the other predictor = 0, which is now its mean). Without centring, $b_1$ would represent the effect of rainfall when phosphorus = 0 (which may be biologically implausible), making the main-effect coefficients uninterpretable.

*b.*

- ✓ $H_0$: $b_3 = 0$ (the effect of rainfall on growth does not depend on soil phosphorus — the two predictors act additively and independently).
- ✓ $H_A$: $b_3 \neq 0$ (the effect of rainfall on growth depends on soil phosphorus, or equivalently, the effect of phosphorus depends on rainfall).
- ✓ Biological meaning of a significant interaction: the growth benefit of additional rainfall changes depending on soil nutrient availability. For example, extra rainfall may strongly increase growth on phosphorus-rich soils (where water is the limiting factor) but have little effect on phosphorus-poor soils (where nutrient limitation caps growth regardless of water).

*c.*

- ✓ $b_0$ = 4.82 cm year$^{-1}$: the predicted growth rate at the **mean rainfall and mean soil phosphorus** (since both predictors are centred). This is the model's baseline — the estimated growth of a site with average environmental conditions.
- ✓ $b_1$ = 0.031 cm year$^{-1}$ per mm rainfall: at mean phosphorus, each additional mm of annual rainfall is associated with a 0.031 cm year$^{-1}$ increase in growth. Rainfall has a small positive effect on growth.
- ✓ $b_2$ = −0.14 cm year$^{-1}$ per mg kg$^{-1}$ phosphorus: at mean rainfall, each additional mg kg$^{-1}$ of soil phosphorus is associated with a 0.14 cm year$^{-1}$ *decrease* in growth. This may seem counterintuitive and could reflect a negative covariation in this landscape (e.g., high-phosphorus soils may be drier or more degraded).
- ✓ $b_3$ = 0.0018: the interaction coefficient. For each unit increase in phosphorus, the slope of rainfall increases by 0.0018. A positive interaction means that the positive effect of rainfall on growth becomes stronger at higher phosphorus levels — i.e., the two resources are synergistic.

*d.*

- ✓ Compare the model **with** and **without** the interaction term using **AIC** (or an F-test / likelihood-ratio test). If $\Delta AIC > 2$ in favour of the model with the interaction, the interaction term provides a meaningful improvement in fit relative to its added complexity.
- ✓ Also examine the **p-value** for $b_3$ in the full model: if $p < 0.05$, there is statistical evidence for a non-zero interaction. However, AIC is preferable as the primary criterion because it penalises model complexity and is not subject to the binary significance threshold.

# Part C: Statistical Output Interpretation (37 marks)

## Question 13 — Welch's *t*-Test Output: Cuttlefish Mantle Length (/12)

The following R output compares mantle length (mm) of *Sepia officinalis* caught at two sites: an inshore estuary and an offshore reef.

```
Welch Two Sample t-test

data: mantle_length by site
t = 3.271, df = 31.84, p-value = 0.002614
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 8.432 35.568

sample estimates:
 mean in group estuary mean in group reef
             142.3              120.0
```

```
Levene's Test for Homogeneity of Variance
        Df  F value   Pr(>F)
group    1   9.412   0.00374 **
      46
```

a. Why was Welch's *t*-test used rather than Student's *t*-test? Support your answer with evidence from the output. **(/ 2)**

b. State the null and alternative hypotheses being tested and interpret the *t*-statistic and *p*-value in plain language. **(/ 3)**

c. Interpret the **95% confidence interval** [8.432, 35.568]. What does it tell you about the precision of the estimated difference? **(/ 3)**

d. The degrees of freedom are reported as 31.84 rather than a whole number. Explain why this occurs in Welch's test and what it implies about the two groups' sample sizes and/or variances. **(/ 2)**

e. Write a one-sentence conclusion reporting the result in the style of a scientific paper. **(/ 2)**

Tip

**Model Answer — Question 13**

*a.*

- ✓ Levene's test is significant ($F(1, 46) = 9.412$, $p = 0.00374$), indicating that the variances differ significantly between the two groups. The homoscedasticity assumption of Student's *t*-test is violated.
- ✓ Welch's *t*-test is appropriate because it does not assume equal variances — it adjusts both the test statistic and the degrees of freedom to account for heteroscedasticity.

*b.*

- ✓ $H_0$: the true difference in mean mantle length between estuary and reef cuttlefish is zero (μ_estuary − μ_reef = 0).
- ✓ $H_A$: the true difference is not equal to zero (μ_estuary ≠ μ_reef).
- ✓ $t = 3.271$, $p = 0.0026$: the observed difference in means (22.3 mm) is 3.271 standard errors from zero. The probability of observing a difference this large or larger by chance, assuming $H_0$ is true, is 0.0026. We reject $H_0$ at $\alpha = 0.05$ and conclude that mean mantle length differs significantly between sites.

*c.*

- ✓ We are 95% confident that the true difference in mean mantle length (estuary − reef) lies between **8.43 mm and 35.57 mm**.
- ✓ The interval is relatively wide (spanning ~27 mm), reflecting moderate imprecision in the estimate. However, since the entire interval is positive (the lower bound exceeds zero), we can be confident the estuary cuttlefish are genuinely longer on average — the direction of the difference is clear even if the exact magnitude is uncertain.

*d.*

- ✓ In Welch's test, the degrees of freedom are calculated using the **Welch-Satterthwaite equation**, which weights each group's contribution by its sample variance and sample size. When the two groups have unequal variances (as indicated here by Levene's test), the formula yields a non-integer value — a fractional reduction from the maximum possible df (N − 2 = 46).
- ✓ The fractional df (31.84 vs. a maximum of 46) implies that the two groups contribute unequally to the pooled uncertainty — likely because one group has a substantially larger variance, which downweights the effective sample size of that group. This results in a more conservative test.

*e.*

- ✓ "Mantle length of *Sepia officinalis* was significantly greater at the inshore estuary (mean = 142.3 mm) than at the offshore reef (mean = 120.0 mm; Welch's *t*-test, $t(31.84) = 3.271$, $p = 0.003$, 95% CI of difference: [8.4, 35.6] mm)."

## Question 14 — Polynomial Regression Output: Zebrafish Growth Rate (/12)

A developmental biologist measures the growth rate (mm day$^{-1}$) of larval zebrafish (*Danio rerio*) at seven temperatures (18, 20, 22, 24, 26, 28, 30°C), with 12 replicate fish per temperature ($n$ = 84 total). She fits three models:

- **M1** (linear): `growth ~ temperature`
- **M2** (quadratic): `growth ~ temperature + I(temperature^2)`
- **M3** (cubic): `growth ~ temperature + I(temperature^2) + I(temperature^3)`

```
Model comparison (AIC):
         M1       M2       M3
AIC  182.41   97.63  100.08

Analysis of Variance (M2 vs M1):
        Df   Sum Sq Mean Sq F value   Pr(>F)
M1       1   38.217
M2       1   29.841  29.841   71.34  < 2e-16 ***

Quadratic model (M2) summary:
lm(formula = growth ~ temperature + I(temperature^2), data = zebrafish)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       -18.432      2.041   -9.03  < 0.001 ***
temperature         1.847      0.183   10.09  < 0.001 ***
I(temperature^2)  -0.0381     0.0041   -9.29  < 0.001 ***

Residual standard error: 0.648 on 81 df
Multiple R-squared: 0.8243, Adj R-squared: 0.8198
F-statistic: 189.4 on 2 and 81 df, p-value: < 2.2e-16
```

a. Using the AIC values, explain which model provides the best balance between fit and complexity. What is the ΔAIC between the best and second-best models, and is the difference meaningful? **(/ 3)**

b. The F-test comparing M2 to M1 is highly significant. What does this test evaluate specifically, and what does the result mean? **(/ 2)**

c. Write the equation of the fitted quadratic model (M2) and use it to predict the growth rate at 24°C. Show your working. **(/ 3)**

d. At what temperature does M2 predict **maximum** growth rate? Show the calculation using calculus or algebra. **(/ 2)**

e. A colleague suggests fitting M3 because $R^2$ will always be higher with more predictors. Critique this reasoning and explain why M3 is not preferred here. **(/ 2)**

Tip

**Model Answer — Question 14**

*a.*

- ✓ **M2 (quadratic)** has the lowest AIC (97.63) and is therefore the best model. The ΔAIC between M2 and M3 is 100.08 − 97.63 = **2.45**; between M2 and M1, ΔAIC = 182.41 − 97.63 = **84.78**.
- ✓ A ΔAIC > 2 is conventionally considered a meaningful difference: M2 is decisively better than M1 (huge ΔAIC), and modestly better than M3 (ΔAIC = 2.45, just above the threshold). M2 is preferred over M3: the cubic term adds complexity without sufficient improvement in fit.

*b.*

- ✓ The F-test comparing M2 to M1 evaluates whether **adding the quadratic term** (temperature²) provides a statistically significant improvement in fit beyond the linear model — i.e., it tests $H_0$: $b_{\text{quadratic}} = 0$.
- ✓ $F(1, 81) = 71.34$, $p < 2 \times 10^{-16}$: the result is highly significant. Adding the quadratic term explains significantly more variance in growth rate than the linear term alone. There is strong evidence for a curved (quadratic) relationship between temperature and growth.

*c.*

- ✓ Equation: growth = −18.432 + 1.847 × temperature − 0.0381 × temperature².
- ✓ At temperature = 24°C: growth = −18.432 + 1.847 × 24 − 0.0381 × 24² = −18.432 + 44.328 − 0.0381 × 576 = −18.432 + 44.328 − 21.946 = **3.95 mm day⁻¹**.

*d.*

- ✓ Maximum occurs where d(growth)/d(temperature) = 0: d/dT [−18.432 + 1.847T − 0.0381T²] = 1.847 − 2 × 0.0381 × T = 0 1.847 = 0.0762 × T T = 1.847 / 0.0762 = **24.2°C**.
- ✓ The model predicts maximum larval growth rate at approximately **24°C**.

*e.*

- ✓ $R^2$ always increases (or stays the same) as more predictors are added, even if the additional terms have no real explanatory value. This is why $R^2$ alone is an unreliable criterion for model selection.
- ✓ AIC penalises model complexity: M3's AIC (100.08) is higher than M2's (97.63) despite having one more parameter, indicating that the cubic term does not improve fit enough to justify its added complexity. Moreover, the cubic term in M3 is likely not statistically significant (*p*-value would be large) and adding it risks **overfitting** — the model would fit noise in the current data rather than capturing a real biological relationship, reducing predictive accuracy for new data.

## Question 15 — Two-Way ANOVA Output with Interaction: Invertebrate Recolonisation (/13)

A marine ecologist performs a field experiment to study how **substrate type** (rock, sand) and **disturbance frequency** (low, medium, high) affect the recolonisation rate (individuals $m^{-2}$ $month^{-1}$) of mobile invertebrates on a rocky shore. Each combination of substrate × disturbance is replicated 10 times ($n$ = 10 per cell, $N$ = 60 total).

```
Two-way Analysis of Variance

Response: recolonisation_rate

                        Df  Sum Sq  Mean Sq  F value   Pr(>F)
substrate                1  1842.4  1842.4    28.71  < 0.001 ***
disturbance              2  3216.8  1608.4    25.07  < 0.001 ***
substrate:disturbance    2  1147.2   573.6     8.94  0.00042 ***
Residuals               54  3465.6    64.2

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

```
Group means (individuals m⁻² month⁻¹):
              Low     Medium    High
Rock          48.3     38.1     21.4
Sand          31.6     28.4     26.9
```

a. Write the full statistical model for this two-way ANOVA including all terms. Define each term. **(/ 3)**

b. Interpret the **interaction effect** using both the ANOVA table and the group means table. Is it valid to interpret the main effects in isolation? Explain. **(/ 4)**

c. Calculate the **total variance** explained by the model (SS_model / SS_total) and interpret this value. **(/ 2)**

d. Conduct a **Bonferroni correction** for the three F-tests reported (substrate, disturbance, interaction). Are all three effects still significant after correction? Show your working. **(/ 2)**

e. Write a **two-sentence conclusion** describing the key ecological findings of this experiment. **(/ 2)**

Tip

**Model Answer — Question 15**

*a.*

- ✓ The full statistical model:

  $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$

  where:

  - $Y_{ijk}$ = recolonisation rate for the $k$th replicate in substrate level $i$ and disturbance level $j$
  - $\mu$ = overall grand mean
  - $\alpha_i$ = effect of substrate type $i$ (rock or sand)
  - $\beta_j$ = effect of disturbance frequency level $j$ (low, medium, or high)
  - $(\alpha\beta)_{ij}$ = interaction effect between substrate $i$ and disturbance $j$
  - $\varepsilon_{ijk} \sim N(0, \sigma^2)$ = residual error, assumed independently and normally distributed with constant variance.

*b.*

- ✓ The interaction term is significant ($F(2, 54) = 8.94$, $p = 0.00042$), indicating that the **effect of disturbance on recolonisation rate differs between rock and sand substrates**.
- ✓ From the group means: on **rock**, recolonisation declines steeply from low (48.3) to high disturbance (21.4) — a decrease of 26.9 individuals $m^{-2}$ $month^{-1}$. On **sand**, recolonisation shows only a slight decline from low (31.6) to high (26.9) — a decrease of only 4.7 individuals $m^{-2}$ $month^{-1}$. The effect of disturbance is thus much stronger on rock than on sand.
- ✓ Because the interaction is significant, the main effects of substrate and disturbance **cannot be interpreted in isolation**: the advantage of rock over sand narrows and even reverses at high disturbance, so a single overall "effect of disturbance" averaged across substrates is misleading.

*c.*

- ✓ SS_model = SS_substrate + SS_disturbance + SS_interaction = 1842.4 + 3216.8 + 1147.2 = 6206.4.
- ✓ SS_total = SS_model + SS_residuals = 6206.4 + 3465.6 = **9672.0**.
- ✓ $\eta^2$ = SS_model / SS_total = 6206.4 / 9672.0 = **0.642** (64.2%). The full model (substrate + disturbance + interaction) explains approximately **64%** of the total variation in invertebrate recolonisation rate, which is a substantial proportion.

*d.*

- ✓ Three tests are conducted, so the Bonferroni-corrected threshold = 0.05 / 3 ≈ **0.0167**.
- ✓ Substrate: $p < 0.001 < 0.0167$ ✓ significant after correction.
- ✓ Disturbance: $p < 0.001 < 0.0167$ ✓ significant after correction.
- ✓ Interaction: $p = 0.00042 < 0.0167$ ✓ significant after correction.
- ✓ All three effects remain statistically significant after the Bonferroni correction.

*e.*

- ✓ "Invertebrate recolonisation rate was significantly affected by substrate type, disturbance frequency, and their interaction ($p < 0.001$ for all terms), with the full model explaining 64% of total variation. Rock substrates supported substantially higher recolonisation at low disturbance, but this advantage diminished at high disturbance levels, where recolonisation rates on rock and sand converged — indicating that rocky-shore invertebrates are

*End of Version 10*

# Bibliography