BCB744 Biostatistics — Theory Test (Version 7)

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 — Simpson's Paradox (/6)

a. Define **Simpson's paradox** in your own words. Provide a brief hypothetical biological example that illustrates the phenomenon. **(/ 3)**

b. How does Simpson's paradox illustrate the importance of **controlling for confounding variables** in statistical analysis? What does it reveal about the danger of combining data across heterogeneous groups without accounting for group structure? **(/ 3)**

Tip

Model Answer — Question 1

a.

- ✓ **Simpson's paradox** occurs when a trend or association observed in several separate subgroups disappears or reverses when the subgroups are combined into a single aggregate dataset. The direction of the relationship "flips" depending on whether a confounding third variable is considered or ignored.
- ✓ **Biological example**: a drug trial tests recovery rates from a bacterial infection in two hospitals. In Hospital A, Drug X has a higher recovery rate than Drug Y (70% vs. 60%). In Hospital B, Drug X also outperforms Drug Y (40% vs. 30%). However, when the data from both hospitals are pooled (because Hospital B treats far more severe cases and received more Drug X patients), Drug Y appears to have a higher overall recovery rate than Drug X — because Hospital B's baseline recovery rate is lower and it contributes disproportionately to Drug X's aggregate statistics.

b.

- ✓ Simpson's paradox demonstrates that an observed association between two variables can be entirely an **artefact of an unmeasured confounding variable** (here, hospital or disease severity) that differentially affects the distribution of patients across groups. Failing to account for this confound leads to incorrect causal conclusions.
- ✓ Aggregating data across heterogeneous groups without adjustment is dangerous because the **composition of each group** (the mix of confounders) can dominate the aggregate result, masking or reversing the true within-group relationship. Stratified analysis, ANCOVA, or inclusion of confounders as covariates are essential safeguards.

Question 2 — Mixed-Effects Models: Concepts (/8)

a. What is a **mixed-effects (hierarchical / multilevel) model**, and in what type of biological study design is one necessary? **(/ 3)**
b. Distinguish between a **fixed effect** and a **random effect** in a mixed-effects model. Give a biological example of each within the same study. **(/ 3)**
c. What is a **random intercept model**? Describe (in words) what the fitted lines would look like for different individuals in such a model, compared to a standard linear regression with no random effect. **(/ 2)**

Tip

Model Answer — Question 2

a.

- ✓ A **mixed-effects model** is a statistical model that contains both fixed effects (parameters representing the effects of specific predictor variables of interest, estimated without assuming they are random) and random effects (parameters representing variation among grouping units sampled from a larger population).
- ✓ It is necessary when observations are **not independent** due to hierarchical or clustered data structure — e.g., when multiple measurements are taken on the same individual over time (repeated measures), when subjects are nested within groups (students within schools, fish within rivers, leaves within plants), or when the same experimental units appear across multiple conditions.

b.

- ✓ A **fixed effect** is a predictor whose specific levels are the focus of inference — conclusions apply to exactly those levels. Example: the effect of three antibiotic concentrations (0, 50, 100 μg mL^{-1}) on bacterial growth — the researcher is interested in *these specific* concentrations and wants to estimate and compare their effects.
- ✓ A **random effect** is a factor whose levels are treated as a random sample from a broader population of possible levels; inference generalises to the full population. Example, in the same study: individual culture flasks, with multiple measurements per flask. The researcher is not interested in specific flask effects but wants to account for the correlation among measurements within the same flask by modelling flask-to-flask variability as a random term.

c.

- ✓ A **random intercept model** includes a random effect for the intercept — each grouping unit (e.g., each individual animal or each study site) is allowed its own baseline level of the response. All units share the **same slope** for the fixed predictor but differ in their **vertical starting point** (intercept).
- ✓ Visually: the fitted lines for different individuals are **parallel** (equal slopes) but **vertically offset** — some individuals have systematically higher responses across all predictor values, others lower. In contrast, a standard linear regression with no random effects produces a single regression line that treats all observations as coming from a homogeneous population, ignoring individual baseline differences.

Question 3 — Regression to the Mean (/5)

a. Define **regression to the mean**. Provide one biological example where this phenomenon could mislead a researcher into concluding that a treatment was effective when it was not. **(/ 3)**

b. How does including a **control group** in an experimental design protect against incorrectly attributing regression to the mean to a treatment effect? **(/ 2)**

Tip

Model Answer — Question 3

a.

- ✓ **Regression to the mean** is the statistical phenomenon whereby extreme values on one measurement occasion tend to be followed by less extreme values on a subsequent measurement, simply due to random variation — not because of any intervention.
- ✓ **Biological example**: a researcher selects 20 animals with the highest parasite loads from a population and treats them with an anti-parasitic drug. On re-measurement, parasite loads are substantially lower. The researcher concludes the drug worked. However, animals selected for extreme high parasite loads were partly selected because of random measurement noise at the time of selection. On re-measurement, the random error component averages out, and their true (lower) parasite loads are revealed — an apparent improvement that would have occurred even without treatment (regression to the mean).

b.

- ✓ A **control group** receives no treatment but is subject to the same selection, measurement, and follow-up protocol as the treatment group. Any regression to the mean affects **both groups equally**.
- ✓ By comparing the treatment group's change to the control group's change (rather than to its own baseline), regression to the mean is cancelled out in the difference. Only changes that are **larger in the treatment group than in the control group** can be attributed to the treatment, providing a valid causal comparison.

Question 4 — Fisher's Exact Test and Proportions (/6)

a. When is **Fisher's exact test** preferred over a chi-square test for analysing a 2×2 contingency table? **(/ 2)**

b. A researcher records whether 200 individual kelp plants show signs of disease at two sites: Site A (80 of 110 diseased) and Site B (40 of 90 diseased). State the null hypothesis for the appropriate test and explain why the test is justified. **(/ 2)**

c. The test returns $p = 0.024$. Interpret this result and calculate the observed disease **prevalence** (proportion) at each site. **(/ 2)**

Tip

Model Answer — Question 4

a.

- ✓ Fisher's exact test is preferred when **expected cell frequencies are small** — specifically, when one or more expected cells fall below 5 (the conventional threshold for the chi-square approximation to be reliable). It computes exact probabilities from the hypergeometric distribution rather than relying on a large-sample approximation.
- ✓ It is always valid (not just for small samples) and is often preferred for 2×2 tables regardless of sample size, as it makes no distributional approximation. It is particularly important for rare events, small case-control studies, or pilot experiments with modest n.

b.

- ✓ H_0: The proportion of diseased kelp plants is equal at Site A and Site B — disease prevalence is independent of site ($\text{proportion}_A = \text{proportion}_B$).
- ✓ The test is a **chi-square test of independence** (or Fisher's exact test if expected cell counts are small). Here, all expected counts are ≥ 5: expected(A, diseased) = 200 × (120/200) × (110/200) ≈ 66, so the chi-square approximation is valid. However, Fisher's exact test is also acceptable as it is conservative and exact.

c.

- ✓ Prevalence at Site A: 80/110 ≈ **0.727** (72.7%); at Site B: 40/90 ≈ **0.444** (44.4%).
- ✓ $p = 0.024 < \alpha = 0.05$: we reject H_0. There is statistically significant evidence that disease prevalence differs between the two sites — Site A has a substantially higher proportion of diseased individuals (~73%) than Site B (~44%).

Question 5 — Effect Size and Practical Significance (/7)

a. What is **Cohen's *d***, and how is it calculated? What values conventionally correspond to small, medium, and large effects? **(/ 3)**

b. A pharmaceutical trial reports $p = 0.002$ for a reduction in bacterial infection rate, but Cohen's $d = 0.08$. What does this combination tell you, and why is the statistically significant result potentially misleading? **(/ 2)**

c. What is the difference between a **confidence interval** and an **effect size** as tools for reporting results, and why should both be reported? **(/ 2)**

Tip

Model Answer — Question 5

a.

- ✓ **Cohen's *d*** is a standardised measure of the difference between two group means, expressed in units of the pooled standard deviation: $d = (\mu_1 - \mu_2) / SD_{pooled}$, where $SD_{pooled} = \sqrt{[(SD_1^2 + SD_2^2) / 2]}$. It is dimensionless and scale-free.
- ✓ Conventional benchmarks (Cohen, 1988): **small** $d = 0.2$, **medium** $d = 0.5$, **large** $d = 0.8$. These are rough guidelines — what constitutes a meaningful effect depends on the biological context.

b.

- ✓ A significant *p*-value (0.002) combined with a tiny effect size ($d = 0.08$, well below 0.2) indicates that the study had a **very large sample size**, which gave it high power to detect even negligible differences. Statistical significance here reflects the precision of the estimate, not the biological or clinical importance of the finding.
- ✓ The significant result is potentially misleading because $d = 0.08$ represents a trivially small effect — the difference in infection rates between drug and control is less than one-tenth of a standard deviation. A statistically significant but practically irrelevant finding can waste resources and mislead clinical or conservation decision-making.

c.

- ✓ An **effect size** (e.g., Cohen's *d*, *r*, η^2) describes the **magnitude** of an effect in standardised or meaningful units, quantifying how large the difference or relationship is regardless of sample size.
- ✓ A **confidence interval** describes the **precision and uncertainty** of the estimated effect — it shows the range of plausible values for the true effect, combining information about both the magnitude and the sampling variability.
- ✓ Both should be reported together: the effect size conveys *how large* the effect is; the CI conveys *how precisely* it is known. A large effect with a wide CI may be uncertain; a small effect with a narrow CI may be real but trivial. Together they provide a complete picture.

Question 6 — Cramér's V and Measures of Association (/6)

a. What is **Cramér's V**, and how is it calculated from a chi-square statistic? What values indicate weak, moderate, and strong association? **(/ 3)**

b. A chi-square test of independence returns $\chi^2(3) = 12.18$ for a 4×2 contingency table with $n = 240$. Calculate Cramér's V and interpret the strength of association. **(/ 2)**

c. Why is Cramér's V a more informative measure of association than the chi-square statistic alone? **(/ 1)**

Tip

Model Answer — Question 6

a.

- ✓ **Cramér's V** = $\sqrt{(\chi^2 / (n \times k))}$, where k = min(rows − 1, columns − 1) is the smaller of the number of rows minus one and the number of columns minus one. It ranges from 0 (no association) to 1 (perfect association), providing a standardised measure of association in contingency tables.
- ✓ Conventional benchmarks (for k = 1, i.e., 2×2 or n×2 tables): **small** V ≈ 0.10, **medium** V ≈ 0.30, **large** V ≈ 0.50. For larger tables (k > 1), the benchmarks are lower (e.g., small ≈ 0.07, medium ≈ 0.21, large ≈ 0.35 for 3-column tables).

b.

- ✓ For a 4×2 table: k = min(4 − 1, 2 − 1) = min(3, 1) = **1**.
- ✓ Cramér's V = $\sqrt{(12.18 / (240 \times 1))} = \sqrt{(0.0508)} \approx$ **0.23**.
- ✓ V = 0.23 indicates a **weak-to-moderate** association between the row and column variables — the association is statistically significant (p = 0.0068) but modest in magnitude. The categorical variables are related, but the relationship explains only a small fraction of the variation in category membership.

c.

- ✓ The chi-square statistic is sensitive to **sample size**: the same underlying effect size will produce a larger χ^2 with a larger n, making χ^2 unsuitable for comparing association strength across studies of different sizes. Cramér's V is **standardised** — it removes the influence of n and table dimensions, allowing the strength of association to be compared across studies and table sizes on a common 0–1 scale.

Question 7 — Robust Statistics and Non-Parametric Alternatives (/7)

a. What makes a statistical method **robust**? Give two examples of robust measures of location and spread used in descriptive statistics. **(/ 3)**

b. The **Wilcoxon signed-rank test** is a non-parametric alternative to the paired t-test. Describe the key steps in performing this test and state the assumption it makes about the data (which the sign test does not require). **(/ 2)**

c. A researcher has 15 pairs of measurements with two extreme outliers. They consider three options: (i) paired t-test on raw data, (ii) paired t-test after log-transformation, (iii) Wilcoxon signed-rank test. Briefly compare these options in terms of validity and power for this dataset. **(/ 2)**

Tip

Model Answer — Question 7

a.

- ✓ A method is **robust** if its validity and performance are not strongly affected by moderate departures from assumptions (e.g., non-normality, presence of outliers, heteroscedasticity). Robust methods give reliable estimates and inferences even when the data deviate from the idealised conditions the method was designed for.
- ✓ Robust measures of location: **median** (not sensitive to extreme values, unlike the mean) and **trimmed mean** (mean calculated after removing the most extreme observations, e.g., the top and bottom 5% or 10%).
- ✓ Robust measures of spread: **interquartile range (IQR)** (based on the central 50% of the data, not affected by outliers) and **median absolute deviation (MAD)** (the median of the absolute deviations from the median).

b.

- ✓ Key steps in the Wilcoxon signed-rank test: (1) Compute the **paired differences** ($d_i = y_{after,i} - y_{before,i}$) for each pair. (2) Discard pairs where $d_i = 0$. (3) **Rank** the absolute values of the differences from smallest to largest, assigning mid-ranks to ties. (4) Assign the **sign** of each original difference to its rank. (5) Compute the test statistic W = sum of positive signed ranks. (6) Compare W to the reference distribution under H_0 (or use a normal approximation for large n).
- ✓ The Wilcoxon signed-rank test assumes that the **distribution of paired differences is symmetric** around the median difference — it does not require normality but does require symmetry. The sign test makes no distributional assumptions at all, but has less power.

c.

- ✓ (i) **Paired *t*-test on raw data**: potentially invalid if the two outliers break the normality assumption of paired differences (especially with only 15 pairs). Outliers will inflate the SD, reducing power.
- ✓ (ii) **Paired *t*-test after log-transformation**: valid if the transformation achieves approximate normality of the differences. May be the most powerful option if the log scale is the appropriate biological scale, but requires that the transformation actually resolves the outlier problem.
- ✓ (iii) **Wilcoxon signed-rank test**: valid without normality, robust to outliers (ranks limit the influence of extreme values), and makes only the symmetry assumption. With 15 pairs and outliers, this is likely the safest option in terms of validity; it may have slightly less power than a valid *t*-test but more power than the *t*-test on data with unresolved outliers.

Question 8 — Generalised Linear Models: Gamma Family (/8)

a. What type of response variable is most appropriately modelled with a **Gamma GLM**? Why is the Gamma distribution more appropriate than the Gaussian for this type of data? **(/ 3)**

b. What is the most commonly used **link function** for a Gamma GLM, and what does it imply about the relationship between the predictor and the response? **(/ 2)**

c. A researcher models **water clarity** (measured as Secchi depth in m, strictly positive continuous variable) as a function of total phosphorus concentration (μg L^{-1}, continuous) and season (Summer vs. Winter; Winter is the reference). Why might a Gamma GLM with log link be preferred over a standard linear model for this response variable? **(/ 3)**

Tip

Model Answer — Question 8

a.

- ✓ A **Gamma GLM** is most appropriate for response variables that are **positive, continuous, and right-skewed** — for example, reaction times, concentrations, rainfall amounts, survival times, or any measurement that cannot take negative values and whose variance tends to increase with the mean.
- ✓ The Gaussian distribution is symmetric and allows negative values, making it a poor fit for strictly positive, skewed data. It also assumes constant variance, whereas many positive continuous measurements show variance that scales with the mean (the coefficient of variation is approximately constant). The Gamma distribution explicitly models this variance-mean relationship ($\text{Var}(Y) \propto \mu^2$) and is defined only for positive values.

b.

- ✓ The most common link function for a Gamma GLM is the **log link**: $\log(\mu) = \eta$, implying that the predictors have a **multiplicative** (exponential) effect on the expected response. On the response scale, a unit increase in a predictor multiplies the expected Secchi depth by e^{β}.
- ✓ The log link ensures that predicted responses are always **positive** (since $e^{\eta} > 0$ for any η), and it is the canonical link for the Gamma family, which makes it the default choice in most statistical software.

c.

- ✓ Secchi depth is **strictly positive** and likely right-skewed (many clear observations, occasional very turbid low-clarity values). A standard linear model can produce **negative predicted Secchi depths** at high phosphorus concentrations, which is biologically impossible.
- ✓ Secchi depth measurements also typically show **variance that increases with the mean** — clearer water (higher Secchi depth) tends to show more variation than turbid water, violating the constant-variance assumption of ordinary linear regression.
- ✓ A Gamma GLM with log link models the expected Secchi depth on a multiplicative scale (consistent with the known inverse relationship between phosphorus and light penetration), uses the appropriate Gamma error distribution for positive continuous data with increasing variance, and constrains all predictions to positive values — better aligning the model structure with the biology and measurement properties of the response.

Question 9 — Intraclass Correlation and Cluster Structure (/8)

a. What is the **intraclass correlation coefficient (ICC)**, and what does it measure in the context of clustered data? **(/ 3)**

b. A researcher measures leaf mass (g) of 80 trees across 8 forest plots (10 trees per plot). The variance components from a one-way random-effects ANOVA are: plot variance = 3.21, residual variance (within plot) = 0.94. Calculate the ICC and interpret its value. **(/ 3)**

c. What does a high ICC (e.g., ICC = 0.80) imply for the **effective sample size** of a study with clustered data? Why is ignoring clustering problematic? **(/ 2)**

Tip

Model Answer — Question 9

a.

- ✓ The **intraclass correlation coefficient (ICC)** measures the **proportion of total variance** in the response that is attributable to differences between clusters (groups), rather than to differences within clusters between individuals. It quantifies how similar observations within the same cluster are relative to observations from different clusters.
- ✓ ICC = $\sigma^2_{\text{between}}$ / ($\sigma^2_{\text{between}}$ + σ^2_{within}), where $\sigma^2_{\text{between}}$ is the variance among cluster means and σ^2_{within} is the variance within clusters. ICC ranges from 0 (no within-cluster resemblance; clustering is irrelevant) to 1 (all observations within a cluster are identical; all variation is between clusters).

b.

- ✓ ICC = plot variance / (plot variance + residual variance) = 3.21 / (3.21 + 0.94) = 3.21 / 4.15 ≈ **0.774**.
- ✓ An ICC of 0.77 means that approximately **77% of the total variation in leaf mass** is attributable to differences between forest plots (i.e., among-plot variation). Only 23% of the variation occurs among trees within the same plot.
- ✓ Biologically: trees within the same plot are substantially more similar in leaf mass to each other than to trees in different plots — strong within-plot resemblance, likely due to shared soil, microclimate, and local species composition effects.

c.

- ✓ A high ICC means that observations within the same cluster are highly similar — they carry less **independent information** than an equal number of truly independent observations. The **effective sample size** (n_{eff}) is substantially smaller than the nominal n: $n_{\text{eff}} \approx n$ / (1 + (cluster size − 1) × ICC). With ICC = 0.80 and clusters of 10, $n_{\text{eff}} \approx 80$ / (1 + 9 × 0.80) = 80 / 8.2 ≈ 10 — much smaller than the nominal 80.
- ✓ Ignoring clustering treats all 80 observations as independent, dramatically underestimating the standard error of treatment effect estimates and leading to inflated test statistics, overly narrow confidence intervals, and an inflated Type I error rate — results will appear far more precise and significant than the data actually support.

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

Question 10 — Nested Design: Leaf Herbivory (/12)

An ecologist measures the percentage of leaf area consumed by herbivores (%) for individual leaves nested within branches, which are in turn nested within trees, across two forest patches. There are 3 trees per patch, 5 branches per tree, and 3 leaves per branch. The first six rows of the dataset are:

	patch	tree	branch	leaf	herbivory_pct
1	1	1	1	1	12.4
2	1	1	1	2	14.1
3	1	1	1	3	11.8
4	1	1	2	1	18.7
5	1	1	2	2	16.3
6	1	1	2	3	19.2

The researcher's question is: *"Is there significant variation in herbivory between the two forest patches?"*

a. Identify the hierarchical levels of the study from broadest to narrowest, and state which level is the **correct experimental unit** for testing patch differences. **(/ 3)**
b. Why would (i) treating each leaf as an independent replicate, and (ii) averaging all leaves within a tree before analysis, both be inappropriate strategies? **(/ 4)**
c. What type of statistical model is most appropriate for this hierarchical data structure? **(/ 3)**
d. What is the correct number of **degrees of freedom** available for testing the patch effect, and why is it so small? **(/ 2)**

Tip

Model Answer — Question 10

a.

- ✓ Hierarchical levels from broadest to narrowest: **Patch** → **Tree** (nested within patch) → **Branch** (nested within tree) → **Leaf** (nested within branch).
- ✓ The correct experimental unit for testing patch differences is the **tree** (or arguably the patch, but with only 2 patches there are no degrees of freedom for a formal patch test without treating trees as replicates of patch). The treatment (patch) is applied at the patch level; trees are the independently sampled units within patches — they are the **replicates** for the patch comparison.
- ✓ With 3 trees per patch and 2 patches, there are 6 trees in total. The patch effect is tested against among-tree-within-patch variation, not against among-leaf or among-branch variation.

b.

- ✓ (i) **Treating each leaf as independent**: the 90 leaves are not independent — leaves on the same branch, branch on the same tree, and tree in the same patch share common environmental conditions and biological histories. This is severe **pseudoreplication**: the effective n is far smaller than 90, inflating the apparent precision and the Type I error rate. Conclusions about patch differences based on 90 pseudoreplicates are invalid.
- ✓ (ii) **Averaging leaves within trees** produces 6 tree-level means (one per tree), which are genuine independent observations. This is appropriate for a simple tree-level analysis. However, it discards all information about within-tree and within-branch variation, which may itself be of biological interest, and it fails to leverage the full hierarchical structure. If averaging is done at the branch level rather than the tree level, the replicates (branches) are still nested within trees and not independent.

c.

- ✓ A **linear mixed-effects model** (nested random effects ANOVA) is most appropriate: `herbivory ~ patch + (1 | patch:tree) + (1 | patch:tree:branch)`. This model includes patch as the fixed effect of interest, and tree-within-patch and branch-within-tree as nested random effects that account for the hierarchical correlation structure. Alternatively, a **nested ANOVA** with patch as the treatment and trees as the random-effect grouping variable can be used.

d.

- ✓ With 2 patches and 3 trees per patch, the patch effect has **df = patches − 1 = 1** for the numerator, tested against df = total trees − patches = 6 − 2 = **4 error degrees of freedom** (the tree-within-patch MS).
- ✓ It is so small because patch is a **between-tree factor** with only 3 tree-replicates per patch (a total of 6 trees). No matter how many leaves or branches are measured, the degrees of freedom for the patch effect are determined by the number of independent units at the level of application of the treatment — the trees. The many sub-samples (branches and leaves) do not add degrees of freedom for this comparison.

Question 11 — Logistic Regression: Coral Bleaching (/13)

A reef ecologist surveys 120 coral colonies at sites spanning a range of SST anomalies (°C above the long-term mean; continuous) and at two depths (Shallow: ≤ 5 m, Deep: > 5 m). Each colony is scored as bleached (1) or healthy (0). The first six rows are:

	colony_id	sst_anomaly	depth	bleached
1	1	0.4	Shallow	0
2	2	0.8	Shallow	1
3	3	1.2	Shallow	1
4	4	0.3	Deep	0
5	5	0.6	Deep	0
6	6	1.1	Deep	1

The research question is: *"Do SST anomaly and depth predict the probability of coral bleaching?"*

a. Why is **logistic regression** appropriate for modelling this response variable? Why is standard linear regression inappropriate? **(/ 3)**

b. State the formal null and alternative hypotheses for the **effect of SST anomaly** on bleaching probability. **(/ 3)**

c. The fitted logistic model returns a coefficient of $\beta = 2.14$ for `sst_anomaly`. Calculate the **odds ratio** and interpret it in plain biological language. **(/ 4)**

d. What does it mean for a logistic regression model to be **well calibrated**? Name one method used to assess calibration. **(/ 3)**

Tip

Model Answer — Question 11

a.

- ✓ The response variable (`bleached`) is **binary** — it takes only the values 0 or 1. Logistic regression is designed specifically for binary outcomes: it models the probability of the event (bleaching) as a function of predictors, constraining predictions to (0, 1) via the logit link and inverse-logit (sigmoid) function.
- ✓ Standard linear regression is inappropriate because it can produce **predicted probabilities outside [0, 1]** (predicting probabilities > 1 or < 0, which is biologically impossible); it assumes **normally distributed residuals** with constant variance, but Bernoulli residuals have variance $p(1 - p)$ that changes with the probability (violating homoscedasticity); and it imposes a linear relationship, whereas the true probability-predictor relationship is S-shaped.

b.

- ✓ H_0: SST anomaly has no effect on the probability of coral bleaching — the logistic regression coefficient for SST anomaly (β_{SST}) = 0. Equivalently, the log-odds of bleaching does not change with SST anomaly.
- ✓ H_A: SST anomaly is associated with the probability of bleaching; $\beta_{SST} \neq 0$. (A directional alternative $\beta_{SST} > 0$ — higher anomalies increase bleaching probability — is biologically justified and acceptable if stated a priori.)
- ✓ The test is conducted using a Wald z-test or a likelihood ratio test comparing models with and without SST anomaly.

c.

- ✓ Odds ratio = $e^{\beta} = e^{2.14} \approx$ **8.50**.
- ✓ For each 1°C increase in SST anomaly above the long-term mean, the **odds of a coral being bleached increase by a factor of approximately 8.5**, holding depth constant.
- ✓ Biologically: corals at sites experiencing even a modest 1°C thermal anomaly have roughly 8.5 times higher odds of bleaching than corals at sites with no anomaly. This extremely large odds ratio is consistent with the well-documented thermal sensitivity of the coral-zooxanthellae symbiosis: bleaching probability rises steeply with even small positive deviations from the thermal tolerance threshold.

d.

- ✓ A logistic regression model is **well calibrated** if the predicted probabilities match the observed frequencies of the outcome across the full range of predicted values — i.e., among all cases assigned a predicted probability of 0.70, approximately 70% should actually be bleached.
- ✓ One method to assess calibration is the **Hosmer-Lemeshow test**: observations are grouped by deciles of predicted probability, and the observed counts in each bin are compared to the model's expected counts using a chi-square statistic. A non-significant result ($p > 0.05$) suggests adequate calibration. **Calibration plots** (predicted probability on the x-axis vs. observed proportion on the y-axis) are a complementary graphical method.

Question 12 — Mixed-Effects Model: Bird Feeding Rates (/12)

A behavioural ecologist records the **feeding rate** (pecks min^{-1}) of 30 individual great tits (*Parus major*) observed on **5 different days** each, during two seasons (Breeding vs. Non-breeding; Non-breeding is the reference). The same 30 birds appear on all 5 days. Air temperature (°C) is recorded on each day as a continuous predictor. The first eight rows of the dataset are:

	bird_id	day	season	temp_C	feeding_rate
1	1	1	Non-breeding	12.4	18.3
2	1	2	Non-breeding	10.8	16.1
3	1	3	Breeding	14.2	24.7
4	1	4	Breeding	15.9	27.3
5	2	1	Non-breeding	12.4	21.8
6	2	2	Non-breeding	10.8	19.4
7	2	3	Breeding	14.2	29.1
8	2	4	Breeding	15.9	31.5

The research question is: *"Do air temperature and season predict the feeding rate of great tits?"*

a. Why would a standard linear regression be **inappropriate** for this dataset? **(/ 3)**
b. Identify the most appropriate statistical model and explain its structure, specifying what is treated as a fixed effect and what as a random effect. **(/ 4)**
c. What is the interpretation of the **random intercept** in this model? What biological variation does it capture? **(/ 3)**
d. How would you extend the model to test whether the **effect of temperature on feeding rate differs between seasons**? **(/ 2)**

Tip

Model Answer — Question 12

a.

- ✓ The data have a **repeated measures / hierarchical structure**: each of the 30 birds contributes 5 observations, and the 150 measurements are therefore not independent. Birds that are naturally high (or low) feeders will contribute systematically high (or low) values across all days, creating **within-bird correlation** that standard regression cannot handle.
- ✓ Ignoring this structure would treat the 150 observations as 150 independent data points, severely **underestimating standard errors**, producing test statistics that are too large, and inflating the Type I error rate — a form of pseudoreplication (treating within-bird measurements as if they came from different birds).

b.

- ✓ A **linear mixed-effects model** (also called a multilevel model) with bird as a random intercept: `lmer(feeding_rate ~ temp_C + season + (1 | bird_id), data = tits)`
- ✓ **Fixed effects**: `temp_C` (continuous) and `season` (categorical, 2 levels) — these are the predictors of primary biological interest, whose effects are estimated and interpreted.
- ✓ **Random effect**: `(1 | bird_id)` — a random intercept for each bird, allowing each individual to have its own baseline feeding rate. This captures between-individual variation in personality, body condition, or territory quality, and appropriately models the correlation among repeated observations from the same bird.

c.

- ✓ The **random intercept** for each bird (bird_id) represents that bird's deviation from the population-level mean intercept — its individual baseline feeding rate, relative to the average bird.
- ✓ Biologically, it captures **individual variation in baseline feeding propensity**: some birds are consistently faster feeders (positive random intercept) and others consistently slower (negative random intercept), independent of temperature and season. This variation may arise from differences in individual personality (bold vs. shy), body mass, territory quality, competitive dominance, or sex — all unmeasured sources of individual-level heterogeneity.

d.

- ✓ To test whether the temperature effect on feeding rate differs between seasons, add a **`temp_C × season` interaction term** to the model: `lmer(feeding_rate ~ temp_C * season + (1 | bird_id), data = tits)`
- ✓ If the interaction term is statistically significant ($p < 0.05$), the slope of temperature on feeding rate differs between breeding and non-breeding seasons — for example, warming may increase feeding rate more strongly during breeding (when energy demands are higher) than during non-breeding. Compare the models with and without the interaction using a likelihood ratio test or AIC.

Part C: Statistical Output Interpretation (37 marks)

Question 13 — Chi-Square Test Output (/12)

An ecologist surveys 270 rockhopper penguins (*Eudyptes chrysocome*) at three breeding colonies (North, Central, South; 90 per colony) and classifies each individual as a First-year, Immature, or Adult breeder. The contingency table and chi-square output are:

```
          First-year  Immature  Adult  Total
North             28        31     31     90
Central           18        26     46     90
South             11        19     60     90
Total             57        76    137    270

    Pearson's Chi-squared test

data:  age_class by colony
X-squared = 21.47, df = 4, p-value = 0.000254
```

a. State the null and alternative hypotheses for this test. **(/ 2)**
b. What does `df = 4` indicate about the contingency table structure? **(/ 2)**
c. Calculate the **expected frequency** for the First-year–North cell. Show your working. **(/ 2)**
d. Interpret the result ($p = 0.000254$) at $\alpha = 0.05$. **(/ 2)**
e. Calculate **Cramér's V** for this table and interpret the strength of association. **(/ 4)**

Tip

Model Answer — Question 13

a.

- ✓ H_0: The distribution of age classes is **independent** of colony — the proportion of first-year, immature, and adult breeders is the same at all three colonies.
- ✓ H_A: Age class distribution is **not independent** of colony — at least one colony has a different age class profile than the others.

b.

- ✓ df = (rows − 1) × (columns − 1) = (3 − 1) × (3 − 1) = 2 × 2 = **4**.
- ✓ This is consistent with a **3 × 3 contingency table** (three colonies × three age classes), as described.

c.

- ✓ Expected frequency = (Row total × Column total) / Grand total = (90 × 57) / 270 = 5130 / 270 = **19.0**.
- ✓ The expected count for First-year breeders at the North colony under independence is 19.0, compared to the observed 28 — substantially more first-years than expected, suggesting over-representation of younger breeders at the northern colony.

d.

- ✓ $p = 0.000254 \ll \alpha = 0.05$: we **strongly reject** H_0. There is very strong statistical evidence that the age class distribution of rockhopper penguins differs significantly among the three colonies.
- ✓ The observed pattern — more first-year and immature breeders in the North, more adults in the South — is far more extreme than would be expected by chance if colony and age class were independent.

e.

- ✓ For a 3 × 3 table: $k = \min(3 - 1, 3 - 1) = \min(2, 2) = \mathbf{2}$.
- ✓ Cramér's V = $\sqrt{(\chi^2 / (n \times k))} = \sqrt{(21.47 / (270 \times 2))} = \sqrt{(21.47 / 540)} = \sqrt{(0.03976)} \approx \mathbf{0.199}$.
- ✓ V ≈ 0.20 indicates a **weak-to-moderate** association for a 3×3 table (benchmarks for $k = 2$: small ≈ 0.07, medium ≈ 0.21, large ≈ 0.35). The association is highly statistically significant but of modest practical magnitude — colony predicts age class membership to a limited degree.
- ✓ Biologically: the gradient from younger breeders in the north to older breeders in the south may reflect colony age (recently established northern colonies attract first-time breeders) or habitat quality gradients influencing age at first breeding and colony fidelity.

Question 14 — Mixed-Effects Model Output (/12)

An ecophysiologist measures leaf-level photosynthetic rate (μmol CO_2 m^{-2} s^{-1}) in 12 individual *Ficus benjamina* plants, each measured at 8 light levels (PAR, μmol photons m^{-2} s^{-1}). A random-intercept mixed-effects model is fitted. The output is:

```
Linear mixed model fit by REML

Formula: photosynthesis ~ PAR + (1 | plant_id)

Fixed effects:
             Estimate Std. Error t value
(Intercept)    1.2340     0.3120    3.96
PAR            0.0214     0.0031    6.90

Random effects:
 Groups    Name        Variance  Std.Dev.
 plant_id  (Intercept)  4.2140    2.053
 Residual               0.8720    0.934

Number of obs: 96, groups: plant_id, 12
REML criterion at convergence: 231.4
```

a. Identify the **fixed effect** and the **random effect** in this model. Why is a random effect for plant_id necessary? **(/ 3)**
b. Interpret the fixed-effect coefficient for PAR (0.0214) in biological terms. **(/ 2)**
c. The random effects table shows plant_id variance = 4.214 and residual variance = 0.872. What does each variance component represent? Which is larger, and what does this tell you about the data? **(/ 4)**
d. Calculate the **intraclass correlation coefficient (ICC)** from these variance components and interpret it in biological terms. **(/ 3)**

Tip

Model Answer — Question 14

a.

- ✓ **Fixed effect**: PAR (photosynthetically active radiation, continuous) — the predictor of primary interest, whose slope is estimated and assumed to represent the population-level relationship between light intensity and photosynthetic rate.
- ✓ **Random effect**: plant_id (individual plant intercepts) — modelling the between-plant variation in baseline photosynthetic capacity. Each plant is allowed its own intercept (baseline photosynthesis at PAR = 0), drawn from a normal distribution with mean zero and variance 4.214.
- ✓ The random effect is necessary because each plant is measured 8 times (once per light level). These repeated measurements from the same plant are **not independent** — plants with inherently high photosynthetic capacity will show elevated rates across all light levels, creating within-plant correlation. The random intercept captures this individual-level variation and prevents pseudoreplication.

b.

- ✓ For each 1 μmol photon m^{-2} s^{-1} increase in PAR, the expected photosynthetic rate increases by approximately **0.0214 μmol CO_2 m^{-2} s^{-1}**, holding plant identity constant.
- ✓ Biologically: a strong positive light-response relationship — higher light availability drives greater carbon fixation. For a 100 μmol photon increase in PAR (a modest increase), the expected photosynthesis increases by 100 × 0.0214 ≈ 2.14 μmol CO_2 m^{-2} s^{-1}, consistent with the initial linear portion of a photosynthesis-light response curve.

c.

- ✓ The **plant_id variance (4.214)** represents **between-plant variation in baseline photosynthetic rate** — the variance in individual plants' intercepts around the population mean intercept. Large plant-to-plant differences in photosynthetic capacity (due to genetics, leaf age, chlorophyll content, or acclimation history) contribute to this term.
- ✓ The **residual variance (0.872)** represents **within-plant variability** — the variance in photosynthetic rate not explained by PAR or by the plant's random intercept (residual measurement-to-measurement variation within the same plant).
- ✓ Plant_id variance (4.214) is **much larger** than residual variance (0.872), indicating that most of the total variation in photosynthesis is **between plants rather than within plants**. Different plants have substantially different baseline photosynthetic capacities, and the within-plant light response (the fixed effect of PAR) is relatively consistent.

d.

- ✓ ICC = plant_id variance / (plant_id variance + residual variance) = 4.214 / (4.214 + 0.872) = 4.214 / 5.086 ≈ **0.829**.
- ✓ An ICC of 0.83 means that approximately **83% of the total variation in photosynthetic rate** is attributable to systematic differences between individual plants. Only 17% of the variation is within-plant (across light levels, beyond the PAR effect).
- ✓ Biologically: individual plant identity is by far the dominant source of variation in photosynthetic rate — knowing which plant is being measured is more informative than knowing the light level. The random intercept model is well-justified and highly necessary; treating each of the 96 measurements as independent would vastly underestimate standard errors and inflate the apparent significance of the PAR effect.

Question 15 — Gamma GLM Output: Water Clarity (/13)

A limnologist models Secchi depth (m; strictly positive, right-skewed) at 65 lakes as a function of total phosphorus concentration (μg L^{-1}, continuous) and land use category (Agriculture vs. Forest; Forest is the reference level). A Gamma GLM with log link is used. The output is:

```
Call:
glm(formula = secchi_depth ~ total_P + land_use,
    family = Gamma(link = "log"), data = lakes)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.1870     0.1240   17.64  < 0.001 ***
total_P             -0.0183     0.0031   -5.90  < 0.001 ***
land_useAgriculture -0.4720     0.1030   -4.58  < 0.001 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.114)

    Null deviance: 18.742  on 64  degrees of freedom
Residual deviance:  7.613  on 62  degrees of freedom
AIC: 143.2
```

a. Why is a Gamma GLM more appropriate than a standard linear model for Secchi depth? **(/ 2)**
b. Write the **fitted equation** for Secchi depth on the log scale and interpret the intercept. **(/ 2)**
c. Interpret the coefficient for `total_P` (−0.0183) on the **response (Secchi depth) scale**. **(/ 3)**
d. Interpret the coefficient for `land_useAgriculture` (−0.4720) as a **multiplicative factor** affecting Secchi depth. **(/ 3)**
e. Calculate the **percentage of null deviance explained** by the model and interpret this value. **(/ 3)**

Tip

Model Answer — Question 15

a.

- ✓ Secchi depth is **strictly positive** (cannot be ≤ 0) and **right-skewed** — a few very clear lakes have very high Secchi depths. A standard linear model can produce **negative predicted Secchi depths** at high phosphorus concentrations and assumes constant variance, whereas Secchi depth data typically show variance increasing with the mean.
- ✓ The Gamma GLM with log link constrains all predictions to positive values, accommodates increasing variance with the mean (Var ∝ μ^2), and models the response on a multiplicative (log) scale — appropriate for a variable spanning a wide positive range.

b.

- ✓ On the log scale: log(Secchi_depth) = 2.187 − 0.0183 × total_P − 0.472 × land_useAgriculture
- ✓ The intercept (2.187) is the **log of the expected Secchi depth for a forest lake with total phosphorus = 0**: back-transformed, $e^{2.187} \approx$ **8.91 m** — the baseline expected clarity for a forest lake. (Note: this is an extrapolation if total_P = 0 is outside the observed range; it is a mathematical baseline, not necessarily a biological one.)

c.

- ✓ On the log scale, a one-unit (μg L^{-1}) increase in total phosphorus reduces log(Secchi depth) by 0.0183.
- ✓ On the **response scale**, the expected Secchi depth is multiplied by $e^{-0.0183} \approx$ **0.982** for each 1 μg L^{-1} increase in total phosphorus — a reduction of approximately **1.8% per 1 μg L^{-1}** increase in phosphorus concentration.
- ✓ For a 10 μg L^{-1} increase in total phosphorus, the expected Secchi depth is multiplied by e^(−0.0183 × 10)^ = $e^{-0.183} \approx$ 0.833 — an 16.7% reduction in water clarity. Biologically: eutrophication (elevated phosphorus) promotes algal growth, reducing light penetration and Secchi depth substantially.

d.

- ✓ The coefficient −0.4720 for land_useAgriculture gives a multiplicative factor of $e^{-0.472} \approx$ **0.624** for agricultural lakes relative to forest lakes of the same total phosphorus concentration.
- ✓ Agricultural lakes have expected Secchi depths approximately **37.6% lower** (or 62.4% of the value) compared to forested lakes with identical phosphorus concentrations. This suggests that agricultural land use impairs water clarity by additional mechanisms beyond phosphorus loading alone — possibly through increased suspended sediment, coloured dissolved organic matter, lake depth, or mixing regime differences.

e.

- ✓ The model explains approximately **59% of the null deviance** in Secchi depth — a moderate-to-strong fit for ecological data, considering that algal suspended matter, lake depth, and mixing regime are not included as predictors.
- ✓ Percentage of null deviance explained = (18.742 − 7.613) / 18.742 × 100 = 11.129 / 18.742 × 100 ≈ **59.4%**.
- ✓ The remaining 41% is attributable to natural variation in Secchi depth and measurement error, lake-specific characteristics not captured by the two predictors.

End of Version 7

Bibliography