BCB744 Biostatistics — Theory Test (Version 5)

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 — Chi-Square Tests (/6)

a. Distinguish between a **chi-square goodness-of-fit test** and a **chi-square test of independence**. For each, state the null hypothesis in general terms. **(/ 3)**

b. What is the minimum expected cell frequency conventionally required before applying a chi-square test? What alternative test should be used when this condition is not met in a 2×2 table? **(/ 2)**

c. Chi-square tests require that observations are **independent**. Give one biological example where this assumption would be violated. **(/ 1)**

Tip

Model Answer — Question 1

a.

- ✓ A **chi-square goodness-of-fit test** compares the observed frequencies of a single categorical variable against expected frequencies derived from a theoretical distribution or prior probabilities. H_0: the observed frequencies are consistent with the theoretical (expected) distribution.
- ✓ A **chi-square test of independence** examines whether two categorical variables in a contingency table are statistically associated. H_0: the two categorical variables are independent — the distribution of one variable does not depend on the level of the other.
- ✓ Key distinction: goodness-of-fit uses one variable compared to a known expectation; the test of independence uses two variables and tests whether their joint distribution equals the product of their marginal distributions.

b.

- ✓ The conventional rule is that all **expected cell frequencies should be ≥ 5** (some sources allow no more than 20% of cells below 5, with a minimum of 1). Below this threshold, the chi-square approximation to the null distribution is unreliable.
- ✓ **Fisher's exact test** is the appropriate alternative for small 2×2 tables — it computes exact probabilities from the hypergeometric distribution rather than relying on the chi-square approximation.

c.

- ✓ Any valid example: recording the same individual animal at multiple trapping occasions and treating each capture as a separate observation; sub-sampling multiple leaves from the same plant and treating each leaf as an independent unit; sampling plots within the same field block without accounting for spatial clustering.

Question 2 — Confidence Intervals (/6)

a. Explain the difference between a **standard error** and a **95% confidence interval**. How is a 95% CI calculated from a standard error (assuming a large sample)? **(/ 3)**

b. Two studies report the same sample mean of 45 g. Study A reports a 95% CI of (38, 52); Study B reports a 95% CI of (43, 47). What can you infer about the relative **sample sizes** of the two studies? **(/ 2)**

c. A researcher states: "Because the 95% CIs for two groups do not overlap, their difference is significant at $\alpha = 0.05$." Is the **converse** of this reasoning correct — i.e., does overlap of two 95% CIs necessarily imply non-significance? Explain. **(/ 1)**

Tip

Model Answer — Question 2

a.

- ✓ The **standard error of the mean** (SE = SD / $\sqrt{n}$) quantifies the precision of the sample mean as an estimate of the population mean — it is the standard deviation of the sampling distribution of the mean across hypothetical repeated samples.
- ✓ A **95% confidence interval** is an interval built from the sample mean that would, across repeated sampling, capture the true population parameter 95% of the time. For large samples: 95% CI ≈ $\bar{x}$ ± 1.96 × SE. For small samples, replace 1.96 with the appropriate *t*-critical value: CI = $\bar{x}$ ± *t*~(df, 0.025)~ × SE.
- ✓ The key distinction: SE is a single number describing sampling variability; the CI is an interval that extends above and below the estimate by a scaled multiple of the SE.

b.

- ✓ Study A has a wider CI (width = 14 g) and Study B has a much narrower CI (width = 4 g). Because CI half-width ∝ SE ∝ $1/\sqrt{n}$, the narrower CI in Study B implies a **substantially larger sample size**. If both studies had the same SD, Study B's *n* would be approximately $(7/2)^2 \approx 12$ times larger than Study A's.

c.

- ✓ **No** — overlapping 95% CIs do not necessarily indicate non-significance. Two individual 95% CIs can overlap by up to about half a CI's half-width while the difference between their means is still statistically significant at $\alpha = 0.05$. The correct approach is to construct a CI for the *difference* between the two means (not compare individual CIs visually) or to perform the appropriate significance test directly.

Question 3 — Binomial Distribution (/7)

a. State the **two conditions** that must hold for a random variable to follow a binomial distribution. **(/ 2)**

b. A shorebird lays exactly **4 eggs** per clutch. The probability of each egg hatching successfully is 0.70, independently of other eggs. Name the distribution that describes the number of eggs hatching per clutch, and state its parameters. **(/ 2)**

c. Calculate the **mean** and **variance** of this distribution. Show the formulae and working. **(/ 2)**

d. Why is a **Poisson distribution** inappropriate for modelling the number of eggs hatching per clutch? **(/ 1)**

Tip

Model Answer — Question 3

a.

- ✓ (i) Each trial has exactly **two possible outcomes** (success or failure), and the probability of success (p) is **constant** and identical for all trials.
- ✓ (ii) The trials are **independent** of one another — the outcome of one egg does not affect the probability of any other egg hatching.

b.

- ✓ The number of eggs hatching follows a **Binomial distribution** with parameters **n = 4** (number of trials, i.e., eggs) and **p = 0.70** (probability of success, i.e., hatching): X ~ Bin(4, 0.70).

c.

- ✓ **Mean**: E(X) = $n \times p = 4 \times 0.70 =$ **2.80** eggs.
- ✓ **Variance**: Var(X) = $n \times p \times (1 - p) = 4 \times 0.70 \times 0.30 =$ **0.84** eggs^2.

d.

- ✓ The Poisson distribution models counts of events over a theoretically unlimited range (0, 1, 2, …, ∞). Here the count is **bounded above by 4** (the fixed clutch size). The Poisson distribution has no upper bound, making it structurally inappropriate. Additionally, Poisson requires mean = variance (equidispersion), but here mean (2.80) ≠ variance (0.84) — the binomial's variance is necessarily smaller than the mean when $p < 1$.

Question 4 — Multiple Testing Corrections (/6)

a. What is the **multiple comparisons problem**? Using the family-wise error rate (FWER), explain with a numerical example how the risk of a false positive accumulates when many tests are performed simultaneously. **(/ 3)**

b. Distinguish between a **Bonferroni correction** and a **Benjamini-Hochberg false discovery rate (FDR) correction**. In what type of biological study would FDR control be preferred over Bonferroni, and why? **(/ 3)**

Tip

Model Answer — Question 4

a.

- ✓ When multiple hypothesis tests are performed on the same dataset, the chance of obtaining at least one false positive by chance alone increases with the number of tests — even if all null hypotheses are true.
- ✓ The **family-wise error rate** (FWER) is the probability of making *at least one* Type I error across all tests: FWER = $1 - (1 - \alpha)^k$. For $k = 10$ tests at $\alpha = 0.05$: FWER = $1 - (0.95)^{10} \approx 0.40$. For $k = 20$: FWER ≈ 0.64 — a 64% chance of at least one spurious "significant" result.
- ✓ This is problematic whenever many comparisons are made simultaneously (e.g., all pairwise group contrasts, multiple endpoints, many gene expression probes) without correction.

b.

- ✓ The **Bonferroni correction** divides α by the number of tests ($\alpha_{adj} = \alpha/k$), strongly controlling the FWER at or below α. It is simple and conservative — but with large k, it is very stringent, greatly reducing power (increasing Type II errors).
- ✓ The **Benjamini-Hochberg FDR correction** controls the expected *proportion* of false discoveries among all rejected hypotheses (not the probability of any single false positive). It is less stringent than Bonferroni and retains more power, at the cost of allowing a small proportion of false positives among the discoveries.
- ✓ FDR is preferred in **large-scale genomic, transcriptomic, or proteomic studies** (e.g., identifying differentially expressed genes among thousands of candidates) where some false positives are tolerable and the goal is to identify a candidate list for follow-up experiments, not to make definitive claims about each individual test.

Question 5 — Logistic Regression (/8)

a. Why is standard linear regression **inappropriate** for modelling a binary response variable (e.g., presence/absence of a pathogen)? **(/ 2)**

b. What does the **logit link function** do to the response variable, and why is it used in logistic regression? **(/ 2)**

c. A logistic regression of coral bleaching probability (bleached = 1, not bleached = 0) on maximum monthly sea surface temperature (°C) returns a coefficient of $\beta = 0.64$ for SST. Interpret this as an **odds ratio**. **(/ 2)**

d. What does a **ROC curve** represent when evaluating a logistic regression model, and what does an AUC of 0.50 indicate? **(/ 2)**

Tip

Model Answer — Question 5

a.

- ✓ Linear regression can produce **predicted probabilities outside [0, 1]**, which is biologically meaningless for a binary response. It also assumes normally distributed residuals with constant variance, but a binary response produces Bernoulli residuals whose variance (= $p(1 - p)$) changes with the predicted probability — violating homoscedasticity structurally.
- ✓ Linear regression also imposes a linear (straight-line) relationship between the predictor and the response, whereas the relationship between a predictor and a probability is inherently S-shaped (bounded and nonlinear at the extremes).

b.

- ✓ The **logit transformation** maps a probability $p \in (0, 1)$ to the log-odds: $\text{logit}(p) = \log(p / (1 - p))$, which is unbounded ($-\infty$ to $+\infty$). This removes the boundary constraint, allowing a linear predictor to operate on an unrestricted scale.
- ✓ It is used as the link function because it ensures that the predicted probability, when back-transformed (via the inverse logit / sigmoid function), always lies in (0, 1) — it connects the linear predictor to the probability scale correctly.

c.

- ✓ Odds ratio = $e^{\beta} = e^{0.64} \approx$ **1.90**. For each 1°C increase in maximum monthly SST, the **odds of a coral being bleached increase by approximately 90%** (i.e., become 1.9 times higher), holding all else constant.
- ✓ Biologically: sites with warmer peak temperatures have substantially higher odds of bleaching, consistent with thermal stress being a primary driver of bleaching events.

d.

- ✓ A **ROC (Receiver Operating Characteristic) curve** plots the true positive rate (sensitivity) against the false positive rate (1 − specificity) across all possible classification thresholds. The **AUC** (area under the curve) summarises the model's ability to discriminate between the two outcomes.
- ✓ An AUC of **0.50** indicates that the model performs no better than random chance — it cannot discriminate between bleached and unbleached corals. An AUC of 1.0 would indicate perfect discrimination; values above 0.70 are generally considered acceptable for a predictive model.

Question 6 — Analysis of Covariance (ANCOVA) (/7)

a. What is an **analysis of covariance (ANCOVA)**, and what statistical purpose does including a continuous covariate serve alongside a categorical predictor? **(/ 3)**

b. A researcher compares mean shell mass (g) of mussels from three tidal zones (low, mid, high) while controlling for shell length (mm) as a covariate. What **key assumption** must be verified before interpreting the adjusted group means? **(/ 2)**
c. What is the essential difference between ANCOVA and **multiple linear regression**, even though both can include continuous and categorical predictors in the same model? **(/ 2)**

Tip

Model Answer — Question 6

a.

- ✓ ANCOVA is a linear model that combines ANOVA (categorical predictor: tidal zone) with linear regression (continuous covariate: shell length) to compare **adjusted group means** — group means estimated at a common, typically the grand-mean, value of the covariate.
- ✓ Including a covariate serves two purposes: (i) it **reduces residual variance** by explaining variation in the response that is attributable to the covariate, thereby increasing statistical power to detect group differences; and (ii) it **controls for confounding** — if groups differ systematically in the covariate (e.g., mussels in the low tidal zone tend to be larger), ANCOVA statistically equalises groups on shell length before comparing shell mass.
- ✓ The result is a group comparison adjusted for the linear effect of the covariate — a fairer comparison of the response at a common body size.

b.

- ✓ The key assumption is **homogeneity of regression slopes** (the parallelism assumption): the slope of the shell length–shell mass relationship must be equal (parallel) across all three tidal zone groups. If slopes differ significantly, the covariate adjustment is inconsistent across groups and the adjusted means cannot be meaningfully compared at a single common value.
- ✓ Tested by fitting a model with a `shell_length × tidal_zone` interaction term: a significant interaction ($p < 0.05$) indicates unequal slopes and a violation of the assumption.

c.

- ✓ Both use the same mathematical framework. The distinction is **conceptual and inferential**: ANCOVA frames the categorical predictor as the **primary treatment** of scientific interest, with the covariate as a nuisance variable being statistically controlled. Multiple regression treats all predictors symmetrically as co-equal candidates for explaining variance in the response, with no single predictor privileged as the "treatment." ANCOVA emphasises adjusted group contrasts; regression emphasises the independent contribution of each predictor.

Question 7 — Model Selection (/7)

a. What is the **Akaike Information Criterion (AIC)**? What does it penalise, and what rule is used to decide when two models differ meaningfully? **(/ 3)**

b. Describe the approaches of **forward selection**, **backward elimination**, and **all-subsets selection** for choosing among candidate predictors. **(/ 2)**

c. Why is it problematic to use stepwise selection and then report p-values from the selected model as if the model structure had been pre-specified? **(/ 2)**

Tip

Model Answer — Question 7

a.

- ✓ AIC = $-2 \times$ log-likelihood + $2k$, where k is the number of estimated parameters. It measures the trade-off between **goodness of fit** and **model complexity**, penalising models that add parameters without sufficient improvement in fit.
- ✓ It discourages overfitting without requiring nested models (unlike likelihood ratio tests). The model with the **lowest AIC** is preferred.
- ✓ Practical rule: ΔAIC < 2 — models are empirically equivalent; ΔAIC 2–7 — moderate support for the lower-AIC model; ΔAIC > 10 — strong support. AICc (corrected AIC) should be used when sample size relative to parameters is small ($n/k < 40$).

b.

- ✓ **Forward selection**: starts with an empty model and sequentially adds the predictor giving the greatest improvement in fit, stopping when no further addition helps.
- ✓ **Backward elimination**: starts with the full model and removes the least significant predictor iteratively, stopping when all remaining predictors meet the retention criterion.
- ✓ **All-subsets selection**: evaluates all 2^p possible models for p candidate predictors and selects the best by AIC, BIC, or adjusted R^2. It avoids path-dependency but is computationally expensive for large p.

c.

- ✓ Stepwise selection is a **data-driven search** across many candidate models: the final model is selected because it fits this particular sample well, not because its variables were theoretically justified beforehand.
- ✓ The p-values in the selected model **do not account for the model-selection process** — the effective number of comparisons made is not reflected in any single p-value, inflating the Type I error rate. Coefficient estimates are also biased towards larger magnitudes (selected partly by chance). AIC-based model averaging or pre-registration of the model structure are more defensible approaches.

Question 8 — Overdispersion and Count Data (/7)

a. What is **overdispersion** in count data? How can you detect it, and what are the consequences for inference if it is ignored in a Poisson model? **(/ 3)**

b. What is the **negative binomial distribution**, and how does its variance structure differ from the Poisson? When is it preferred? **(/ 2)**

c. A parasitologist models tick counts per deer host using a Poisson GLM. The residual deviance is 418.3 on 98 degrees of freedom. What does this suggest, and what should the researcher do? **(/ 2)**

Tip

Model Answer — Question 8

a.

- ✓ **Overdispersion** occurs when the observed variance in count data **exceeds the variance expected** under the assumed model. For a Poisson model, the expected variance equals the mean; overdispersion means Var(Y) > E(Y).
- ✓ Detection: compare the residual deviance to the residual degrees of freedom in a fitted Poisson GLM. If the deviance/df ratio (the dispersion parameter) is substantially greater than 1 (e.g., > 2), overdispersion is likely. Formal tests (e.g., the `dispersiontest()` in R's **AER** package) can also be used.
- ✓ If ignored: standard errors are **underestimated**, z-statistics are inflated, and p-values are too small — leading to an inflated Type I error rate and false claims of significant effects.

b.

- ✓ The **negative binomial distribution** is a discrete probability distribution for count data that models the number of events when the underlying rate varies among individuals (i.e., there is extra-Poisson heterogeneity). Its variance = $\mu + \mu^2/k$, where k is the dispersion parameter. As $k \longrightarrow \infty$, the negative binomial approaches the Poisson.
- ✓ Unlike the Poisson (variance = mean), the negative binomial has **variance > mean** — it accommodates overdispersion. It is preferred over Poisson when count data are overdispersed (residual deviance $\gg$ residual df), which is common in ecological data where some individuals carry far more parasites, seeds, or pathogens than others.

c.

- ✓ Residual deviance / df = 418.3 / 98 $\approx$ **4.27** — far greater than 1. This strongly suggests that the Poisson model is **severely overdispersed**: the data show far more variability in tick counts than a Poisson distribution predicts.
- ✓ The researcher should refit the model using a **negative binomial GLM** (or quasi-Poisson with variance $\propto \mu$ as a simpler alternative), which explicitly allows variance to exceed the mean. This will produce valid standard errors and p-values.

Question 9 — Blocking in Experimental Design (/7)

a. What is a **randomised complete block design (RCBD)**? Why is blocking used, and what does it achieve statistically? **(/ 3)**

b. Give one ecological or biological example where blocking would be appropriate, and describe what the block would represent. **(/ 2)**
c. What is the statistical consequence of **failing to account for blocks** in the analysis (i.e., fitting a one-way ANOVA without a block term) when blocks explain substantial variance? **(/ 2)**

Tip

Model Answer — Question 9

a.

- ✓ In a **randomised complete block design**, experimental units are grouped into **blocks** (groups of units that are more similar to one another than to units in other blocks). Within each block, all treatment levels are randomly assigned — each block contains one complete replicate of all treatments.
- ✓ Blocking is used to **control for a known source of extraneous variation** (the blocking factor) that is not of primary interest. By grouping similar units together and ensuring all treatments appear in each block, between-block differences are removed from the error term.
- ✓ Statistically, blocking **reduces the residual variance** and increases the precision of treatment comparisons — it is the equivalent of a paired design extended to more than two treatments.

b.

- ✓ **Example**: comparing the effect of three fertiliser types on grass growth across six fields. Each field (block) differs in soil type, drainage, and history. Within each field, three plots are randomly assigned one fertiliser type each. The block is the **field** — it absorbs between-field variation in background soil fertility, ensuring that fertiliser comparisons are made within homogeneous field conditions.

c.

- ✓ If blocks are omitted from the analysis and their variation is substantial, the **between-block variance is absorbed into the residual (error) term**, inflating the residual mean square.
- ✓ This increases the denominator of the F-ratio, **reducing statistical power** to detect treatment effects — real treatment differences may be masked by the uncontrolled block-to-block noise. Conversely, if blocks are positively correlated with the treatment assignment (by accident), ignoring them can also bias the F-test upward.

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

Question 10 — Habitat Association of Reef Fish (/12)

A marine ecologist surveys the microhabitat use of three reef fish species (Parrotfish, Wrasse, Damselfish) across three habitat types (Coral reef, Seagrass bed, Sandy bottom). For each species, 60 individuals were observed and the habitat in which each individual was found is recorded. The observed counts are:

	Coral_reef	Seagrass	Sandy_bottom
Parrotfish	38	14	8
Wrasse	22	21	17
Damselfish	15	30	15

The research aim is: *"To determine whether habitat use differs significantly among the three reef fish species."*

a. State the formal null and alternative hypotheses for this analysis. **(/ 3)**
b. Identify the statistical test and give **two specific reasons** for your choice, with reference to the nature of the variables. **(/ 3)**
c. The test returns $\chi^2(4) = 18.43$, $p = 0.001$. Interpret this result fully, including reference to the degrees of freedom. **(/ 3)**
d. What additional step would you take to determine **which specific species-habitat combinations** contribute most to the significant result? Describe the quantity you would examine and how you would interpret it. **(/ 3)**

Tip

Model Answer — Question 10

a.

- ✓ H_0: Habitat use is **independent** of species — the proportion of individuals found in each habitat type is the same across all three fish species.
- ✓ H_A: Habitat use is **not independent** of species — at least one species has a different distribution across habitat types compared to the others.
- ✓ The alternative is non-directional: we do not predict in advance which species-habitat combination will depart from independence.

b.

- ✓ **Chi-square test of independence** (Pearson's χ^2).
- ✓ Reason 1: Both variables — species identity and habitat type — are **categorical** (nominal). The chi-square test of independence is the appropriate test for detecting association between two categorical variables summarised in a contingency table.
- ✓ Reason 2: Each of the 180 observations (60 per species) is **independent** — each individual fish is classified into exactly one species × habitat cell, with no repeated measurement or clustering of fish.

c.

- ✓ df = (number of species − 1) × (number of habitats − 1) = (3 − 1) × (3 − 1) = **4** — consistent with the 3×3 contingency table structure.
- ✓ $\chi^2(4) = 18.43$, $p = 0.001 < \alpha = 0.05$: we **reject** H_0. There is strong statistical evidence that habitat use differs significantly among the three reef fish species.
- ✓ The probability of observing a χ^2 statistic of 18.43 or more extreme, given that species identity and habitat use are independent, is only 0.1% — the observed distribution of fish across habitats departs substantially from the expectation under independence.

d.

- ✓ Examine **standardised (Pearson) residuals** for each cell: residual = (Observed − Expected) / √Expected. Cells with |residual| > 2 (approximately equivalent to $p < 0.05$ for a single cell) contribute disproportionately to the overall χ^2.
- ✓ For example, Parrotfish appear strongly over-represented on coral reef (observed 38 vs. expected ≈ 20 if independent) and under-represented on sandy bottom (observed 8 vs. expected ≈ 20), generating large positive and negative residuals respectively.
- ✓ This analysis is analogous to post-hoc testing in ANOVA: the omnibus χ^2 identifies *that* a significant association exists; residual analysis pinpoints *where* in the table the association is concentrated.

Question 11 — Mussel Shell Thickness Across Shore Types (ANCOVA) (/13)

A marine biologist measures shell thickness (mm) in mussels (*Mytilus galloprovincialis*) from three shoreline types (Exposed, Semi-exposed, Sheltered), also recording shell length (mm) as a continuous covariate. The first six rows of the dataset are:

	mussel_id	shore_type	shell_length_mm	thickness_mm
1	1	Exposed	52.1	3.84
2	2	Exposed	48.7	3.61
3	3	Semi-exposed	54.2	3.47
4	4	Semi-exposed	51.8	3.29
5	5	Sheltered	49.3	2.95
6	6	Sheltered	53.6	3.02

The research question is: *"Does shell thickness differ among shore types, after statistically controlling for shell length?"*

a. State formal null and alternative hypotheses appropriate for this ANCOVA. **(/ 3)**
b. Why is it necessary to include shell length as a covariate? What would be the risk of comparing raw (unadjusted) group means? **(/ 3)**
c. State the **key assumption** of ANCOVA that must be verified before interpreting the adjusted means. Describe how you would test it and what outcome would indicate a violation. **(/ 4)**
d. If the assumption in (c) is violated, describe **two alternative approaches** the researcher could use. **(/ 3)**

Tip

Model Answer — Question 11

a.

- ✓ H_0: After adjusting for shell length, the mean shell thickness does not differ among shore types (adjusted μ_{Exposed} = adjusted $\mu_{\text{Semi-exposed}}$ = adjusted $\mu_{\text{Sheltered}}$).
- ✓ H_A: After adjusting for shell length, at least one shore type has a mean shell thickness that differs from the others.
- ✓ The hypotheses explicitly reference the covariate-adjusted means — failing to mention the adjustment would be an incomplete statement of the ANCOVA hypothesis.

b.

- ✓ Shell thickness scales with body size — larger mussels have thicker shells. If mussels from different shore types differ systematically in shell length (e.g., wave-exposed mussels are smaller due to physical disturbance or differential growth), then unadjusted mean thickness differences among shore types would **confound the shore-type effect with body size effects**.
- ✓ Ignoring shell length risks attributing a size-driven difference in thickness to shore type, producing a **biased or spurious treatment effect**. ANCOVA statistically removes the linear effect of shell length and estimates the shore-type effect at a common shell length.

c.

- ✓ The critical assumption is **homogeneity of regression slopes** (parallelism): the slope of the shell length vs. shell thickness relationship must be equal across all shore type groups. If it differs, the covariate adjustment is not uniform and the adjusted means at a single common shell length are not comparable across groups.
- ✓ Test: fit a model including the interaction term `shell_length_mm × shore_type` (i.e., `lm(thickness_mm ~ shell_length_mm * shore_type, data = mussels)`). If the interaction term is statistically significant ($p < 0.05$), the slopes are not equal — the assumption is violated.
- ✓ Graphically, a violation appears as non-parallel regression lines for the three groups when thickness is plotted against shell length, with markedly different slopes (one group's relationship much steeper or shallower than others).

d.

- ✓ Option 1: **Stratified analysis with simple slopes** — report the relationship between shell length and thickness separately for each shore type, and describe how the shore-type effect on thickness varies across the range of shell lengths (Johnson-Neyman technique).
- ✓ Option 2: Fit a **multiple regression with the interaction explicitly included** (`thickness ~ shell_length + shore_type + shell_length:shore_type`) and use this model to obtain predicted thickness at specific, biologically meaningful shell lengths for each shore type, rather than a single overall adjusted mean. This acknowledges and quantifies the heterogeneous slopes rather than forcing parallelism.

Question 12 — Disease Risk in Bottlenose Dolphins (/12)

A marine mammal biologist surveys 90 bottlenose dolphins (*Tursiops truncatus*) at six coastal sites, classifying each individual as clinically healthy (0) or showing signs of lobomycosis skin disease (1). At each site, the researcher records mean water temperature (°C) and proximity to commercial fishing vessels (Near vs. Far; Far is the reference). The first six rows of the dataset are:

```
  dolphin_id  site  diseased  water_temp_C  fishing
1          1     A         0          18.2    Far
2          2     A         1          19.1    Far
3          3     B         0          21.4   Near
4          4     B         1          22.8   Near
5          5     C         1          24.3   Near
6          6     C         0          17.9    Far
```

The research question is: *"Do water temperature and proximity to fishing vessels predict the probability of lobomycosis infection in bottlenose dolphins?"*

a. State formal null and alternative hypotheses for the effect of water temperature on infection probability. **(/ 2)**
b. Identify the most appropriate statistical test and give **three reasons** for your choice, with reference to the response variable and the predictor types. **(/ 4)**
c. If the fitted model returns a coefficient of $\beta = 0.31$ for `water_temp_C`, calculate the **odds ratio** and interpret it biologically. **(/ 3)**
d. What concern would arise from the fact that multiple dolphins were sampled from the same site? How could this be addressed in the analysis? **(/ 3)**

Tip

Model Answer — Question 12

a.

- ✓ H_0: Water temperature has no effect on the probability of lobomycosis infection; the logistic regression coefficient for water temperature (β_{temp}) = 0.
- ✓ H_A: Water temperature is associated with the probability of infection; $\beta_{temp} \neq 0$. (A directional alternative — $\beta_{temp} > 0$, i.e., warmer water increases infection risk — is also acceptable if justified a priori by biological reasoning about fungal growth rates.)

b.

- ✓ **Logistic regression** (binomial GLM with logit link).
- ✓ Reason 1: The **response variable is binary** (0 = healthy, 1 = diseased). Logistic regression is designed for binary outcomes; standard linear regression would produce predicted probabilities outside [0, 1] and violate residual normality.
- ✓ Reason 2: There are **two predictor variables of different types** — one continuous (water_temp_C) and one categorical (fishing: Near vs. Far) — and the model must accommodate both simultaneously. Logistic regression handles mixed predictor types naturally.
- ✓ Reason 3: The goal is to estimate the **probability of disease** as a function of environmental conditions, and to quantify the independent contribution of each predictor via coefficients and odds ratios — this is the core inferential purpose of logistic regression.

c.

- ✓ Odds ratio = $e^{\beta} = e^{0.31} \approx$ **1.36**. For each 1°C increase in water temperature, the **odds of a dolphin being infected with lobomycosis increase by approximately 36%**, holding proximity to fishing vessels constant.
- ✓ Biologically: warmer coastal waters may promote the growth of *Loboa loboi* (or related fungi) or suppress the dolphin's immune response, making infection more likely in warmer areas or seasons. The effect is moderate — a 10°C difference would be associated with odds $\approx 1.36^{10} \approx 19$ times higher, a substantial increase.

d.

- ✓ Multiple dolphins sampled from the same site are likely **not fully independent** — site-level factors (local pollution, prey availability, habitat quality) create shared risk among dolphins at the same site, violating the independence assumption of standard logistic regression.
- ✓ This is a **clustered / hierarchical data structure**. It could be addressed by: (i) fitting a **mixed-effects logistic regression** with site as a random intercept (`glmer(diseased ~ water_temp_C + fishing + (1 | site), family = binomial)`), which explicitly models within-site correlation; or (ii) using **cluster-robust standard errors** in the logistic model, which inflate standard errors to account for non-independence without adding random effects.

Part C: Statistical Output Interpretation (37 marks)

Question 13 — Logistic Regression Output (/12)

An epidemiologist models the probability that a seabird tests positive for avian influenza (1 = positive, 0 = negative) as a function of log-transformed colony size (continuous) and geographic region (North vs. South; South is the reference level). The output is:

```
Call:
glm(formula = infected ~ log_colony_size + region,
    family = binomial, data = seabirds)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -4.2130     0.7821   -5.39  < 0.001 ***
log_colony_size   0.8840     0.1930    4.58  < 0.001 ***
regionNorth       1.1250     0.3420    3.29   0.0010 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 342.1  on 249  degrees of freedom
Residual deviance: 289.4  on 247  degrees of freedom
AIC: 295.4
```

a. Write the **linear predictor (log-odds) equation** for this model. **(/ 2)**

b. Interpret the coefficient for `log_colony_size` (0.8840) as an **odds ratio**, in plain biological language. **(/ 3)**

c. Interpret the coefficient for `regionNorth` (1.1250) in plain biological language. **(/ 3)**

d. The null deviance dropped from 342.1 to 289.4 after adding two predictors. What does this tell you about model fit? **(/ 2)**

e. Why is a **z-statistic** (not a *t*-statistic) reported for each coefficient in this logistic regression? **(/ 2)**

Tip

Model Answer — Question 13

a.

- ✓ log-odds(infected) = −4.213 + 0.884 × log_colony_size + 1.125 × regionNorth
- ✓ (where regionNorth = 1 for northern sites, 0 for southern; accept values rounded to 2–3 d.p.)

b.

- ✓ Odds ratio = $e^{0.884} \approx$ **2.42**. For each one-unit increase in log colony size (i.e., a 10-fold increase in raw colony size), the **odds of infection increase by a factor of approximately 2.42**, holding region constant.
- ✓ Biologically: birds in larger colonies face substantially higher infection odds — consistent with density-dependent transmission, where crowded aggregations facilitate direct contact and airborne pathogen spread.
- ✓ The effect is highly significant ($p < 0.001$), indicating that colony size is a robust predictor of infection risk independent of geographic region.

c.

- ✓ The coefficient 1.125 for regionNorth gives odds ratio = $e^{1.125} \approx$ **3.08**. Northern seabirds have approximately **3.1 times higher odds** of testing positive for avian influenza compared to southern seabirds from colonies of the same size.
- ✓ Biologically: a meaningful geographic gradient exists in infection risk — northern regions face a substantially higher prevalence or transmission rate of avian influenza, independent of colony size. Possible explanations include: overlap with infected migratory waterfowl flyways, greater exposure to arctic breeding shorebirds, or differences in environmental persistence of the virus.

d.

- ✓ The deviance dropped by 342.1 − 289.4 = **52.7** units when two predictors were added (2 df). Under the null model, this follows a χ^2 distribution: $\chi^2(2) = 52.7$, $p \ll 0.001$ — the two predictors together **substantially and significantly improve** model fit beyond the intercept-only baseline.
- ✓ A rough analogue to R^2 for this GLM: 1 − (289.4 / 342.1) ≈ **0.154** — the two predictors explain approximately 15% of null deviance.

e.

- ✓ In logistic regression (a GLM estimated by maximum likelihood), coefficient estimates follow an **asymptotic normal distribution** for large samples, so the Wald test statistic = Estimate / SE is compared to the standard normal (Z) distribution — hence z, not *t*.
- ✓ The *t*-statistic in ordinary least squares regression uses the *t* distribution because the residual variance is estimated from data, adding additional uncertainty. In a binomial GLM the dispersion parameter is fixed at 1 (not estimated from residuals), so the asymptotic normal approximation applies and z is the appropriate test statistic.

Question 14 — Chi-Square Test Output (/12)

A conservation biologist traps 240 hedgehogs at four habitat types (Urban garden, Farmland, Woodland, Road verge; 60 per habitat) and classifies each as Juvenile or Adult. The contingency table and chi-square test output are:

```
               Juvenile  Adult  Total
Urban garden       22      38     60
Farmland           35      25     60
Woodland           28      32     60
Road verge         18      42     60
Total             103     137    240

    Pearson's Chi-squared test

data:  age_class by habitat_type
X-squared = 12.183, df = 3, p-value = 0.0068
```

a. State the null and alternative hypotheses tested by this chi-square analysis. **(/ 2)**
b. What does `df = 3` tell you about the structure of the contingency table? **(/ 2)**
c. Calculate the **expected frequency** for the Juvenile-Urban garden cell. Show your working. **(/ 2)**
d. Interpret the p-value = 0.0068 at $\alpha = 0.05$. State your conclusion clearly. **(/ 2)**
e. A colleague suggests reporting **Cramér's V** for this table. What does Cramér's V measure, and why is it a useful complement to the chi-square test? **(/ 4)**

Tip

Model Answer — Question 14

a.

- ✓ H_0: The age class distribution of hedgehogs is **independent** of habitat type — the proportion of juveniles (and adults) is the same across all four habitats.
- ✓ H_A: Age class distribution is **not independent** of habitat type — at least one habitat has a different juvenile:adult ratio compared to the others.

b.

- ✓ df = (number of rows − 1) × (number of columns − 1) = (4 − 1) × (2 − 1) = 3 × 1 = **3**.
- ✓ This is consistent with a **4 × 2 contingency table** (four habitat types × two age classes) as described.

c.

- ✓ Expected frequency = (Row total × Column total) / Grand total = (60 × 103) / 240 = 6180 / 240 = **25.75**.
- ✓ The expected count for Juvenile-Urban garden is 25.75, compared to the observed 22 — fewer juveniles than expected under independence.

d.

- ✓ $p = 0.0068 < \alpha = 0.05$: we **reject** H_0. There is statistically significant evidence that the juvenile:adult ratio of hedgehogs differs across habitat types.
- ✓ The probability of observing a χ^2 statistic of 12.183 or more extreme by chance, if habitat type and age class are truly independent, is only 0.68% — a convincing departure from independence.

e.

- ✓ **Cramér's V** = $\sqrt{(\chi^2 / (n \times \min(r - 1, c - 1)))}$, where r is the number of rows and c is the number of columns. It measures the **strength of association** between the two categorical variables on a scale from 0 (no association) to 1 (perfect association), standardised for table size and sample size.
- ✓ Here: Cramér's V = $\sqrt{(12.183 / (240 \times 1))} = \sqrt{0.0508} \approx$ **0.23** — a weak-to-moderate association.
- ✓ It is a useful complement because the chi-square statistic alone is sensitive to sample size — with a very large n, even trivially small associations become highly significant ($p \ll 0.001$). Cramér's V separates **effect size** from statistical significance, telling the biologist whether the habitat-age association is practically meaningful (large V) or merely a detectable but biologically negligible pattern (small V).
- ✓ Cramér's V provides a comparable, standardised measure of association across studies with different sample sizes and table dimensions.

Question 15 — ANCOVA Output: Lizard Metabolic Rate (/13)

A physiologist measures resting metabolic rate (RMR, mL O_2 hr^{-1}) of *Lacerta agilis* (sand lizard), comparing males and females while controlling for body mass (g) as a covariate. The `lm()` output is:

```
Call:
lm(formula = RMR ~ body_mass + sex, data = lizards)

Residuals:
    Min      1Q  Median      3Q     Max
 -2.814  -0.891   0.043   0.874   3.127

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.1430     0.4820    4.45  < 0.001 ***
body_mass      0.3870     0.0412    9.39  < 0.001 ***
sexMale        1.8250     0.3140    5.81  < 0.001 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.423 on 57 degrees of freedom
Multiple R-squared: 0.8312,  Adjusted R-squared: 0.8255
F-statistic: 140.2 on 2 and 57 DF,  p-value: < 2.2e-16
```

a. Write the fitted regression equations separately for (i) **female** lizards (reference level) and (ii) **male** lizards. **(/ 3)**
b. Interpret the coefficient for `body_mass` (0.3870) in biological terms. **(/ 2)**
c. Interpret the coefficient for `sexMale` (1.8250) in the ANCOVA context. What does it represent geometrically? **(/ 3)**
d. What does the overall F-statistic ($F(2, 57) = 140.2$, $p < 2.2\times10^{-16}$) test? **(/ 2)**
e. What important assumption of this ANCOVA is **not** tested in the output shown, and how would you test it? **(/ 3)**

Tip

Model Answer — Question 15

a.

- ✓ **(i) Female lizards** (sexMale = 0): $\widehat{RMR} = 2.143 + 0.387 \times body\_mass$
- ✓✓ **(ii) Male lizards** (sexMale = 1): $\widehat{RMR} = (2.143 + 1.825) + 0.387 \times body\_mass = 3.968 + 0.387 \times body\_mass$
- ✓ Both equations have the **same slope** (0.387 mL O_2 hr^{-1} per gram) — consistent with the ANCOVA assumption of homogeneity of regression slopes; they differ only in intercept.

b.

- ✓ For each additional 1 g of body mass, resting metabolic rate is predicted to increase by approximately **0.39 mL O_2 hr^{-1}**, holding sex constant. Larger lizards have a higher absolute metabolic demand — a positive allometric relationship between body mass and resting metabolism.
- ✓ This effect is highly significant ($t = 9.39$, $p < 0.001$), confirming that body mass is a strong predictor of RMR — its inclusion as a covariate appropriately accounts for this major source of variation before comparing sexes.

c.

- ✓ The coefficient 1.825 for sexMale represents the **difference in adjusted mean RMR** between males and females at the same body mass: males have an RMR approximately 1.83 mL O_2 hr^{-1} **higher** than females of equal body mass.
- ✓ Geometrically, this is the **vertical distance between the two parallel regression lines** — the male line lies 1.825 units above the female line at every value of body mass. This parallel offset (same slope, different intercept) is the defining geometric signature of an ANCOVA model with no interaction term.
- ✓ Biologically: males have a higher mass-adjusted metabolic rate than females, possibly related to higher activity levels, reproductive investment, or hormonal differences.

d.

- ✓ The overall *F*-test evaluates whether **the model as a whole** (body mass + sex together) explains significantly more variance in RMR than the intercept-only null model — i.e., whether at least one predictor has a non-zero coefficient.
- ✓ $F(2, 57) = 140.2$, $p < 2.2 \times 10^{-16}$: we strongly reject the null model. The combination of body mass and sex explains a highly significant proportion of variance in lizard RMR; neither predictor is irrelevant.

e.

- ✓ The critical assumption not tested here is **homogeneity of regression slopes** (parallelism) — the assumption that the slope of body mass on RMR is equal for males and females. If the slopes differ, the fixed-intercept ANCOVA model mis-specifies the relationship.
- ✓ To test it: fit a model including the interaction term `body_mass:sex` — i.e., `lm(RMR ~ body_mass * sex, data = lizards)`. Compare this to the model without the interaction using an *F*-test (ANOVA table) or AIC. A significant interaction ($p < 0.05$) indicates unequal slopes and violation of the parallelism assumption — in which case the ANCOVA-adjusted means should not be compared directly.

End of Version 5

Bibliography