BIO2010 Data Science for Biologists

MONASH UNIVERSITY · BIOLOGICAL SCIENCES · UNDERGRADUATE

Full guide + ask Sia about BIO2010

SIDE 1/2 CLOSED-BOOK MEMORY SHEET · what must be in your head

SEMESTER 2 · 2026

definitions → decisions → diagnostics → bounded claims

0 · Exam map

COURSE FACTS · W13

P 10 | Q 6 | Assignments 34 | Exam 50

ASSESSMENT	WEIGHT
Pre-class preparation and participation	10%
Four topic quizzes	6%
Three assignments	34%
Theory and concept examination	50%

CLOSED BOOK: NOTHING PRINTED OR WRITTEN COMES INTO THIS EXAM. *BIO2010 Week 13 exam-prep workshop slides (undated; current Moodle capture) (PDFRAW_Week_13_workshop.txt)*

The final is **2 hours 10 minutes** in the current evidence. Revise by retrieval: say each definition, decision rule, assumption and boundary without looking, then apply it to a changed case.

1 · Question → evidence

C1

Research question = focused query naming **response, explanatory factor, population and comparison**. Add the observational unit and scale before choosing a method.

Testable hypothesis predicts an observable pattern and can lose support when that pattern is absent. Null = reference pattern; alternative = informative direction or structure.

research question — testable hypothesis — reproducible workflow

Question fixes the object; hypothesis risks contradiction; workflow preserves evidence.

RECITE THE CHAIN

```
raw file → documented transformation/exclusion → distribution → model → diagnostics → interpretation
```

CHECK → *A precise hypothesis cannot rescue convenience sampling, dependence or outcome measurement after labels are known.*

1A · Retrieval boxes

C1 · FINAL LOGIC

SAY FROM MEMORY	MUST CONTAIN
Question	Response; explanatory factor; population; comparison; observational unit
Hypothesis	Predicted observable pattern; direction or structure; pattern that counts against it
Workflow	Untouched input; recorded exclusions; transformations; graph; model; diagnostics
Conclusion	Effect/pattern; uncertainty; population/range; evidence that would revise it

2 · Biological data

C2

Observational unit = smallest independent entity carrying response and explanatory information. Write what **one row** means before filtering, pivoting, joining or summarising.

observational unit — tidy data — data transformation

Tidy structure preserves, but never creates, the independent biological unit.

RULE	EXAM TEST
Variable → column	One measured property?
Observation → row	Which entity and occasion?
Unit type → table	Are plot, plant and leaf mixed?
Join checked key	Did distinct-unit counts change?

Types constrain operations: character, factor, integer and continuous values are not interchangeable. Converting a category to a number does not create a quantitative scale. Ordered factors need biological order.

Wrangle audit: retain raw columns; create cleaned fields; count rows and distinct plants/plots before and after duplication/discard risks. Missing ≠ zero ≠ below detection ≠ never scheduled.

REPEATED MEASURES

```
five dates may produce five rows per plant; they do not create five independent plants.
```

BOUNDARY → *Tidy structure does not establish independence, correct measurement error or justify excluding an inconvenient value.*

2A · Wrangle decisions

WORKED C2

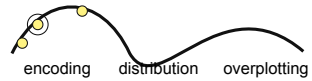
OPERATION	BEFORE	AFTER/CHECK
Pivot dates long	Define plant + occasion	IDs travel; one plant-day per row
Join plot metadata	Check key uniqueness	Distinct plant/plot counts unchanged
Filter records	State biological rule	Count removals; preserve source field
Group summary	Name grouping level	State denominator: measurements, plant means or plot means
Transform response	Retain raw value	Interpret back on biological scale

A second sensor file with duplicated timestamps is a key problem, not permission to drop duplicates. Make duplication visible with counts; decide the intended key from the data-generating process. A cleaning log records original field, new field, rule, affected records and biological reason.

3 · Visual evidence

C3

QUESTION	FIRST DISPLAY
Continuous response across categories	Individuals/distribution; reveal assignment level
Two continuous variables	Points before fitted line
Hidden block/group	Colour or facet; retain denominator
Overlap	Transparency/jitter; do not invent x-variation



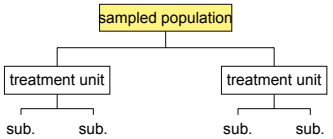
A graph exposes scale, spread, replication and overlap; it cannot create causality.

Position is usually compared more accurately than area or colour intensity. Record axis limits, transformations and smoothing. A caption states what is shown and which summary/interval is used—not untested significance.

TRAP → *A compelling plot cannot distinguish treatment effect from confounding, selection bias or pseudoreplication.*

4 · Design hierarchy

C4



Replication comes from independently assigned units, not every measured branch.

Random allocation balances unmeasured causes in expectation for treatment comparison. **Random sampling** supports generalisation. **Blocking** groups comparable units; **blinding** reduces differential treatment or measurement.

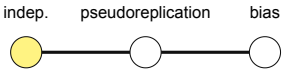
4A · Design controls

C3–C4

FEATURE	CONTROLS/SUPPORTS	DOES NOT ENSURE
Random sampling	Representativeness in expectation	Causal treatment contrast
Random allocation	Comparable treatment groups	Population generalisation
Blocking	Known nuisance variation	Removal of all confounding
Blinding	Differential behaviour/measurement	Independence or random sampling
Replication	Information about treatment variation	Correction of systematic bias

5 · Independence, error, bias

C4



Dependence duplicates information; bias displaces estimates systematically.

PATTERN	MEANING	RESPONSE
Random error	Scatter	Independent replication
Systematic bias	Directional displacement	Design/measurement control
Pseudoreplication	Subsamples as replicates	Respect allocation hierarchy

Experimental unit = entity independently assigned treatment. If watering is assigned to a bench, bench—not pot or leaf—is the replicate. A nested model represents lower-level variation but cannot manufacture bench replication.

CLAIM BOUNDARY

```
generalise only to conditions represented by the treatment units and sampling frame.
```

LIMIT → *Significance cannot repair confounding, a biased sensor or one treatment unit per condition.*

6 · Spread ≠ uncertainty

C5

standard deviation — standard error — confidence interval

SD describes observed spread; SE and CI concern an estimated statistic.

SAMPLE SUMMARIES

$\bar{x} = \Sigma x_i / n$

$s^2 = \Sigma(x_i - \bar{x})^2/(n-1)$ · $s = \sqrt{s^2}$

$SE(\bar{x}) = s/\sqrt{n}$, only when n counts independent units

Mean/SD are sensitive to extremes; median/IQR resist more. A confidence interval is neither the range containing most organisms nor a probability that this fixed parameter lies in this interval.

6A · Choose the summary

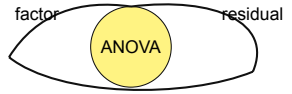
C5

QUANTITY	QUESTION ANSWERED	CLASSIC ERROR
Mean	Arithmetic centre	Ignoring skew/extremes
Median	Middle ordered value	Treating as mean
SD	Spread of observations	Calling it uncertainty in mean
SE	Sampling variability of statistic	Using leaves as independent n
CI	Procedure range for target	Calling it range of organisms

7 · One factor: ANOVA

C6

Model response as grand mean + group effect + residual. Omnibus null: all group means equal in the model. F compares between-group structure with within-group residual variation.



Factor defines groups; F tests structure; residuals challenge fit.

EVIDENCE	SUPPORTS
Small omnibus p	Common-means model fits poorly
Planned contrast	Specified biological comparison
Multiplicity-aware post-hoc	Narrow pairwise claims
Residual pattern	Model-mismatch review

Not licensed: "every pair differs", "large biological importance" or "factor caused response". Observed habitat differences are conditional, not causal.

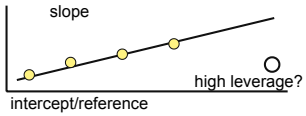
8 · Regression

C7

LINEAR MODEL

$\hat{y} = \beta_0 + \beta_1 x$ · $e = y - \hat{y}$

β_1: response units per predictor unit, within observed range



Leverage is unusual x-position; influence is change to the fitted result.

Residual-versus-fitted: curvature/changing spread. Normal-quantile: error distribution for small-sample inference. Investigate extreme points and refit for sensitivity; never delete solely because the result changes.

8A · Model decoder

C6–C7

PROMPT	MODEL MEANING	REPORT
Groups differ?	Factor indicators vs common mean	F evidence + planned contrast
Continuous relation?	Mean response changes with x	Slope units + observed range
Prediction at x?	Fitted expectation $\beta_0+\beta_1 x$	Mean prediction, not certainty
Unequal spread?	Residual variance changes	Diagnostic consequence/adjustment

9 · Multiple variables

C8

Conditional coefficient = fitted change for one predictor while included predictors are held fixed. Confounders, mediators and colliders alter the target differently; "more variables" is not automatically better.

conditional coefficient — interaction — chi-square test

Conditional effects, changing slopes and categorical association answer distinct questions.

INTERACTION

$y = \beta_0+\beta_1 x+\beta_2 z+\beta_3 xz$

$z=0$: slope β_1 · $z=1$: slope $\beta_1+\beta_3$

With sex coded 0/1, β_1 is the reference-sex slope; β_3 is slope change. β_2 compares sexes at x=0 unless x is centred. Reference changes relabel coefficients, not fitted values.

CATEGORICAL ASSOCIATION

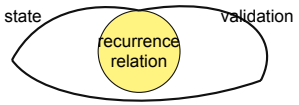
E_{ij}=(row total×column total)/grand total

$X^2=\Sigma(O-E)^2/E$

Row and column percentages answer different questions. Inspect cell contributions and pair χ^2 with a readable proportion difference. Expected counts come from independence, not an even split.

10 · Process models

C9



A state is updated by a rule; evidence must test the resulting behaviour.

POPULATION BALANCE

$n' = n(1+b)(1-d)+m$

state, births, deaths and migration share the stated time step

Verification: code implements equation. **Validation:** behaviour is adequate for purpose/system. Inspect trajectories; vary one defensible input; compare with observations not used for tuning.

BOUNDARY → *Not a forecast beyond parameter range, time step, population closure and measurement assumptions.*

10A · Choose/test the structure

C8–C9

STRUCTURE	CORE QUESTION	DECISIVE CHECK
Additive multiple model	Conditional change, common across groups?	Residuals + coefficient target
Interaction model	Does slope/difference vary by group?	β_3 and fitted group lines

BIO2010 Data Science for Biologists

SELF-TEST DRILLS · ANSWER PATHS · JUDGEMENT CARDS · CLASSIC TRAPS

Full guide + ask Sia about BIO2010

SIDE 2/2 close the sheet · solve aloud · expose the first failed link

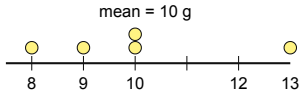
4 WORKED DRILLS

if you see X → do Y · then state the boundary

11 · Drill: SD → SE

WORKED C5

Q. Independent plot means are 8, 9, 10, 10 and 13 g. Compute $\bar{x}$, sample variance, SD and SE; interpret the denominator.



Five independent plot means—not every leaf—define n=5.

ANSWER PATH

$\bar{x}=(8+9+10+10+13)/5=10$ g
deviations: -2, -1, 0, 0, 3
$\Sigma\text{deviation}^2=14$
$s^2=14/(5-1)=3.5$ g²
$s=\sqrt{3.5}\approx1.87$ g
$SE=1.87/\sqrt{5}\approx0.84$ g

Interpret: SD is heterogeneity among plot means. SE is estimated sampling variability of their mean, provided plots are independent and representative. Report units and plot-level denominator.

Changed case: add 30 g: mean and SD move strongly; median/IQR resist. Decide whether error, rare biology or model mismatch—do not delete on sight.

TRAP → *A narrow interval can be precisely wrong under biased sampling, displaced measurement or ignored clustering.*

11A · Uncertainty self-test

ANSWER BEFORE REVEAL

PROMPT	CORRECT RETRIEVAL
What changes if every value gains 5 g?	Mean/median +5; SD, SE and IQR unchanged
What changes if units g→mg?	Centre/spread/SE multiply by 1000; interpretation changes units
Why divide variance by n−1?	Course sample-variance rule after estimating $\bar{x}$
Why not use leaf count for n?	Leaves are subsamples when plots supply independence
What defeats a narrow CI?	Biased sample, systematic error, ignored clustering/model failure

Which is true? "SD is uncertainty in the mean" is false; SD describes between-observation dispersion. "SE is population variability" is false; it estimates repeated-sample variability of a statistic. "This 95% interval has 95% probability of containing the fixed parameter" is not the course definition; coverage belongs to the procedure under assumptions.

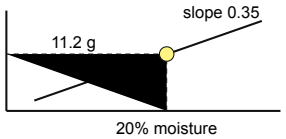
ORAL ANSWER

name statistic → name units and independent n → show arithmetic → distinguish variation from uncertainty → state sampling/design assumption.

12 · Drill: fitted line

WORKED C7

Q. Eight independent plots span 10–30% soil moisture. biomass = 4.2+0.35×moisture. Interpret the slope and predict at 20%.



Substitution gives a fitted mean, not a guarantee for one plant.

ANSWER PATH

$\hat{y}(20)=4.2+0.35\times20=11.2$ g
Within 10–30%, +1 percentage point moisture ↔ +0.35 g fitted mean biomass.

Do not say 0.35 g per 1% relative increase. Do not interpret β_0 biologically when zero lies outside the observed range.

DIAGNOSTIC	JUDGEMENT
Curvature	Linear mean inadequate
Fanning	Variance changes
Extreme x	Check leverage/sensitivity
One-point slope	Investigate; report dependence

Scale change: proportions = percentages/100, so slope becomes 35 g per proportion unit; fitted values and relation stay unchanged.

BOUNDARY → *No extrapolation; observational slope is not a treatment effect.*

12A · Regression retrieval

C7 CHECKS

TERM/CHECK	SAY PRECISELY
β_0	Fitted mean when x=0; meaningful only at sensible zero
β_1	Fitted mean change per one x-unit, within modelled range
Residual e	Observed y minus fitted $\hat{y}$
Leverage	Potential from unusual predictor position
Influence	Actual change to fitted result when point matters
Sensitivity refit	Diagnose dependence; do not silently delete

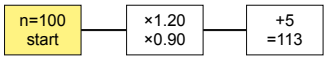
Which is true? A point far in y but ordinary in x may have a large residual without high leverage. A far-x point on the fitted trend can have high leverage but little influence. "Significant slope" does not license extrapolation, causation or prediction of an individual plant.

Diagnostic answer: name the expected residual pattern; describe the observed departure; identify threatened feature (linear mean, stable variance or inference); propose investigation/sensitivity; narrow the claim.

13 · Drill: recurrence

WORKED C9

Q. n=100, b=0.20, d=0.10 and m=5 per step. Calculate n′. Then increase d to 0.25 and judge the direction.



Keep process order and time units attached to each parameter.

BASELINE

$n'=100\times1.20\times0.90+5=113$
if rates stay constant, next step starts at n=113.

CHANGED DEATH RATE

$n'=100\times1.20\times0.75+5=95$
one defensible parameter change reverses growth.

Path: define state/step → map flows → calculate → state constant-rate/closure assumptions → vary one input → identify changed conclusion → compare with withheld observations.

CHECK	QUESTION
Units	Do b,d,m match step?
Verification	Code implements equation?
Validation	Behaviour matches evidence/purpose?
Sensitivity	Which input reverses claim?

CHANGED MECHANISM → *Density-dependent migration needs a stated function, units and evidence before altering the recurrence.*

13A · Model stress test

C9 JUDGEMENT

QUESTION	REQUIRED ANSWER
What is state?	Quantity summarising system at a stated time
What is flow?	Birth/death/migration contribution with units per step
Why this order?	Multiplication represents the stated biological process
Verification evidence?	Hand calculation/known case matches code update
Validation evidence?	Behaviour compared with intended use and observations
Forecast boundary?	Parameter range, step, closure and measurement assumptions

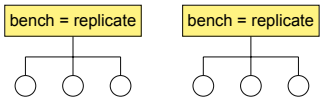
Changed mechanism: density-dependent migration changes the equation, not merely confidence. State function/units, predict direction and name needed observations.

Train/evaluate separation: calibrating parameters and judging success on the same values overstates adequacy. Compare predictions with withheld observations. A smooth or stable trajectory shows numerical behaviour, not that the mechanism is biologically right.

14 · Drill: true replicate

WORKED C4

Q. Watering is assigned to benches; each bench has six pots; three leaves are measured per pot. What is n? What changes if watering is randomised to pots?



Count independent allocation units; measurement abundance is not replication.

Answer: bench supplies replication. Pots/leaves are subsamples. Use several benches per regime; randomise regime to benches; block/rotate location when appropriate.

Changed case: pot becomes experimental unit when treatment is independently randomised to pots; shared bench may remain a blocking/random effect.

SHORT-ANSWER FRAME

allocation level → experimental unit → subsamples/repeats → dependence → supported population → limit

MARKER TEST → *More leaves sharpen each bench response; they do not add treatment replication.*

15 · If X → do Y

C2–C8

SEE	DO
Repeated IDs after join	Test key; count distinct units
Right-skewed response	Raw + justified transform
Several groups	Factor → F → contrast
Two continuous	Scatter → line → residuals
Both categorical	Table → E → χ^2 → proportions

15A · Design/wrangle drill

CHANGED CASES

Sensor join: duplicated plant_id + timestamp is a key problem. Count multiplicity; decide whether sensor_id belongs in the key; join only after defining the biological record per row.

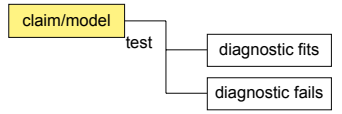
CHANGED CONDITION	WHAT MOVES FIRST?
Treatment bench→pot	Experimental unit and dependence model
Distance→habitat category	Explanatory representation, graph, model
Height→damaged-leaf proportion	Response scale, geometry, uncertainty
Add duplicated timestamps	Join key/record definition—not biology
One habitat variance ×4	Residual diagnostic/inferential adequacy

Greenhouse path: bench allocation → bench replicate → nested pots/leaves → several benches per regime → randomise/block → standardise measurement → limit generalisation.

16 · Which is true?

CLASSIC TRAPS

FALSE SHORTCUT	TRUE CONTRAST
Five dates = five plants	Repeated rows do not change replicate.
Numeric category = quantitative	Conversion does not create magnitude.
Small p: every group differs	F rejects only common-means model.
High leverage = influential	Influence means fit changes.
CI contains most organisms	CI concerns a target estimate.
More predictors = less bias	Mediators/colliders alter target.
Stable code = valid biology	Verification ≠ validation.
Graph proves treatment effect	Pattern cannot remove confounding.



A diagnostic must be capable of narrowing or reversing the preferred claim.

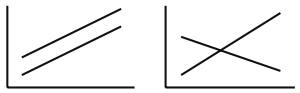
CONDITIONAL LANGUAGE EARNS MARKS

- Within the observed 10–30% range…
- Provided plots are independent…
- Conditional on included predictors…
- For this sampling frame…

BEST ENDING → *Name the observation, design feature or changed parameter that forces revision.*

17 · Interaction + χ^2

C8



Parallel lines: common slope; non-parallel: effect changes by group.

With interaction, neither related main-effect coefficient is universal. Reconstruct fitted values at relevant x/categories. For χ^2, inspect expected counts/cell contributions; sparse-category merging must be scientific, not result-driven.

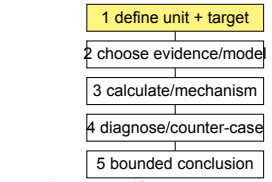
17A · Model contrasts

C6–C8

CONTRAST	DECISION RULE
ANOVA vs pairwise	Omnibus first; contrasts answer named biology
Main vs interaction	With β_3, main terms are reference-specific
Marginal vs conditional	Included predictors change target
χ^2 vs effect size	Test discrepancy; proportions explain pattern

18 · Short-answer engine

ALL CHAPTERS



Five-mark answers need five reasoning moves, not five descriptions.

Answer the verb. Give unit, evidence, relationship/calculation, counter-evidence and a decision inside population, range or system boundary.

CALCULATION

inputs/units → equation → arithmetic → output/interpretation → assumption

COMPARE + DIAGNOSE

common target → decisive difference → consequence; expected pattern → departure → repair → revised claim

SELF-MARK → *Underline response unit, circle independent n, box output, star claim boundary.*

19 · Exam-morning plan

RETRIEVAL ONLY



One short loop per topic: recall, mechanism, transfer, boundary.

1. Recite nine triads without notes.
2. Write SD/SE, regression, interaction, E, χ^2 and recurrence formulas.
3. Name independent unit and one invalidating condition.
4. Redo four drills with numbers/units unaided.
5. Correct the first failed link only.

NINE TRIADS

question → hypothesis → workflow
unit → tidy data → transformation
encoding → distribution → overplotting
independence → pseudoreplication → bias
SD → SE → CI
factor → ANOVA → residual
slope → intercept → leverage
conditional coefficient → interaction → χ^2
state → recurrence → validation