

Experimental Design Midterm

Instructor answer key | Fall 2026

For the seven-page student exam, 2026-midterm.pdf, created October 2, 2026. Question numbering is preserved. Accept equivalent wording and sensible rounding; the exam does not assign point values.

Opening question: What is a p-value?

Assuming the null hypothesis and the statistical model's assumptions are true, the p-value is the probability of obtaining a test statistic at least as extreme as the observed statistic, in the direction or directions specified by the alternative hypothesis.

It is not the probability that the null hypothesis is true, the probability that the result happened "by chance," or a measure of biological importance.

Dataset 1 | Mycorrhizae and plant growth

1. Biological null hypothesis

Mycorrhizal inoculation does not change population mean seedling height after six weeks: the two treatment means are equal.

Accept "inoculation does not increase mean height" for a directional test; formally, its null includes effects of zero or less.

2. Graph and visual checks

Make a side-by-side dotplot of all seedling heights, with treatment on the x-axis and height (cm) on the y-axis. A boxplot with the individual points is also appropriate. Examine differences in center, within-group spread, overlap, skew, and unusual observations.

With only eight plants per group, showing the individual observations is especially useful. Independence comes from the design, not from the graph.

3. Appropriate test

Use an independent two-sample t-test, preferably Welch's test. The plants are separately randomized experimental units, not matched pairs. Welch's test does not require equal variances; independence and approximately normal errors without severe outliers matter here.

Accept a pooled two-sample t-test with a justified equal-variance assumption. A one-sided increase alternative is defensible if selected before examining the results.

4. Biological interpretation

Inoculated seedlings averaged 21.94 cm, compared with 19.78 cm for controls. The estimated increase was **2.16 cm** after six weeks (95% CI: 1.26 to 3.07 cm; two-sided Welch $p = 0.000162$). Random assignment supports an effect of inoculation under these experimental conditions; whether this increase is practically important requires biological context.

Statistical evidence and biological importance

Dataset 2 | A very significant result that may not matter

5. Evidence that growth differs

Yes. The reported $p = 0.0000023$ is below 0.05, and the 95% CI excludes zero. Mean growth was higher in the supplement group by an estimated 0.040 g (95% CI: 0.023 to 0.057 g).

6. Evidence of a biologically important change

Biological importance cannot be determined from these results alone. The estimated increase is 0.040 g, about 2.2% of the control mean. Its importance depends on the biological context or a meaningful-effect threshold, neither of which is specified.

The heading suggests a small practical effect, but the stem does not establish a threshold. Credit students who identify this missing information. Random assignment is also not stated, so a causal conclusion would require additional design information.

7. Which summary is better?

B is the intended better reporting approach because it gives direction, magnitude, statistical support, and biological context. "Highly significant" in A gives no effect size and does not establish importance.

However, B overstates the supplied evidence: its claim that 0.04 g is much smaller than a biologically important difference requires a threshold that the question never provides. A fully supported summary is: "Mean growth was 0.040 g higher with the supplement (95% CI: 0.023 to 0.057 g). The difference was statistically well supported, but its biological importance cannot be determined from the information provided."

Give credit for choosing B with a sound explanation, and for explicitly noting that neither statement is fully satisfactory as written.

Dataset 3 | Potentially important but not statistically significant

8. Why "the treatment has no effect" is too strong

Failing to reject a zero effect does not demonstrate that the effect is zero. The estimated effect is +1.5 eggs, and the 95% CI includes both zero and increases large enough to matter biologically. At the 0.05 significance level, the result is inconclusive rather than evidence of no effect.

9. Information supplied by the confidence interval

The interval shows effect magnitude and uncertainty: effects from a slight decrease of 0.2 eggs to an increase of 3.2 eggs remain compatible with the data under the model. It includes zero, positive effects smaller than the one-egg importance threshold, and biologically important increases. The study therefore does not establish either no effect or a meaningful benefit.

Do not require a statement that there is a 95% probability the fixed true effect lies in this particular interval. That is not the standard frequentist interpretation.

Infection, habitat, and geographic variation

Dataset 4 | Frog infection and habitat

10. Variables and types

Habitat: categorical, with forest and urban categories. **Infection status:** categorical, with infected and not-infected categories. Both are binary categorical variables.

11. Expected infected forest frogs

Under independence, expected count = (forest total × infected total) / grand total = $(100 \times 66) / 200 = 33$ frogs. The common expected infection proportion is $66/200 = 0.33$.

12. Biological interpretation

Infection was more common in sampled forest frogs (42%) than urban frogs (24%), an **18 percentage-point difference**. The chi-square test supports an association between habitat and infection status (Yates-corrected chi-square = 6.536, df = 1, p = 0.0106). This observational comparison does not show that habitat caused the difference.

If a student independently computes an uncorrected Pearson test, $p \approx 0.00679$ is also correct for that stated method; the substantive conclusion is unchanged.

Dataset 5 | Latitude and clutch size

13. Predictor and response

Predictor: latitude in degrees north. **Response:** population mean clutch size, measured in eggs per clutch. Both are quantitative variables.

14. Interpretation of the slope

For each additional degree north in latitude, the fitted model predicts an increase of approximately **0.107 eggs in mean clutch size**. This is an association across populations, not proof of a causal effect of latitude.

15. Predicted difference between 30°N and 40°N

The predicted difference is slope × change in latitude = $0.107 \times (40 - 30) = 1.07$ eggs. Mean clutch size is predicted to be about 1.07 eggs larger at 40°N.

Accept approximately 1.072 eggs when the more precise slope, 0.1072, is used. The intercept cancels when calculating this difference.

Reading regression output

Dataset 6 | Complete clutch-size regression output

16. Meaning of $R^2 = 0.943$

The fitted linear relationship with latitude accounts for **94.3% of the observed variation in population mean clutch size** around its overall sample mean. This is an in-sample measure of fit, not the percentage caused by latitude or the percentage of predictions that are correct.

17. Does the small p-value establish causation?

No. It provides strong evidence against a zero slope under the regression assumptions. Latitude was not randomly assigned, and other variables may covary with it. Statistical significance alone cannot identify the mechanism responsible for the relationship.

18. Plausible covarying variable

For example, **breeding-season length** can vary with latitude and influence reproductive scheduling and clutch size. Temperature, seasonality, food availability, predation, habitat, or population/species differences are also reasonable answers when linked plausibly to both geography and clutch size.

These are possible explanations, not mechanisms established by the data.

Dataset 8 | Significant regression, weak prediction

19. Statistical evidence of a relationship

Yes. The supplied $p = 0.000007$ supports a nonzero slope at the 0.05 level. The estimated association is positive, but the small p-value does not establish a large or useful relationship.

20. Interpretation of the wing-length coefficient

An additional **1 mm of wing length** is associated with an increase of **0.018 eggs in predicted mean lifetime fecundity**. Fractional eggs are appropriate for a modeled mean even though individual egg counts are integers.

21. Contrast the p-value with R^2

The p-value measures evidence against a zero slope under the model. **$R^2 = 0.0016$ means only 0.16%** of the observed variation in fecundity is accounted for by the linear relationship with wing length. With 12,000 observations, a tiny association can be detected precisely while explaining very little variation. The fitted line improves very little over predicting the sample mean alone; out-of-sample performance would need separate assessment.

Repeated measures and integrated interpretation

Dataset 9 | Independence and repeated measurements

22. Why the two samples are not independent

The same mice are measured before and after treatment, so observations form matched pairs and measurements within a mouse may be correlated. There are **30 subjects contributing 30 paired changes**, not 60 unrelated subjects.

23. Appropriate design and test

Use a **paired t-test** to test whether the population mean within-mouse change is zero: `t.test(after, before, paired = TRUE)`. Equivalently, apply a one-sample t-test to `after - before`. The relevant normality assumption concerns the differences, and pairs must be independent across mice.

An appropriate repeated-measures model is also acceptable. Without an untreated comparison group, the design tests change over time but cannot separate treatment effects from aging, growth, or other time effects.

Dataset 10 | P-values and uncertainty

24. Most defensible statement

C. At the 0.05 level, the experiment does not reject a zero effect ($p = 0.061$). The estimate is +4.2%, but the interval from -0.3% to +8.7% permits a slight decrease, no effect, and potentially important increases. A and B are too definite; D incorrectly treats the p-value as the probability that the null is true.

Code comprehension

25. Interpretation of -0.63

For each one-unit increase in temperature, the fitted model predicts a **decrease of 0.63 units in mean beetle body mass**. The prompt does not specify units, so grams or degrees Celsius should not be required. The coefficient describes association unless the study design supports a causal interpretation.

Integrated interpretation

26. Example four-sentence Results paragraph

Mean response was 51.8 in Treatment B and 51.2 in Treatment A, with 4,000 observations per treatment. The estimated B – A difference was 0.6 units (95% CI: 0.4 to 0.8 units). The data provided strong statistical evidence against a zero difference ($p < 0.00001$). However, the entire confidence interval was well below the 5-unit threshold for biological importance, indicating that this statistically supported difference was biologically trivial by the stated criterion.

Credit direction, magnitude, uncertainty, statistical evidence, and comparison with the explicit 5-unit threshold; exact wording is not required.

Instructor checks and grading notes

Numerical checks from the raw data

Calculation	Verified result
Dataset 1: group summaries	Control: $n = 8$, mean 19.775 cm, SD 0.9067 cm. Mycorrhizae: $n = 8$, mean 21.9375 cm, SD 0.7689 cm.
Dataset 1: two-sided Welch test	Mycorrhizae – control = 2.1625 cm. $t = 5.1448$; $df = 13.636$; $p = 0.0001619$. 95% CI: 1.2587 to 3.0663 cm.
Dataset 1: acceptable alternatives	Preselected one-sided increase: $p = 0.00008093$. One-sided 95% interval: 1.4208 cm to infinity. Pooled two-sided test: $df = 14$; $p = 0.0001489$.
Dataset 4: expected counts	Forest: 33 infected, 67 not infected. Urban: 33 infected, 67 not infected.
Datasets 5–6: raw-data regression	Intercept 0.678788; slope 0.107168. $R^2 = 0.9426$; adjusted $R^2 = 0.9369$. Slope $p = 0.0000001567$. Residual standard error ≈ 0.1999 on 10 df.

Corrections and fair grading

Dataset 2, Questions 6–7: No biological-importance threshold is provided. Do not require students to conclude that 0.040 g is trivial; credit a well-reasoned objection to the final clause of summary B.

Dataset 6: The printed residual standard error is 0.211; recomputing from Dataset 5 gives approximately 0.1999. The supplied coefficients and R^2 agree with the data after rounding. This discrepancy does not affect Questions 16–18. Do not penalize students for correctly interpreting the supplied output.

Dataset 8: The displayed $t = 4.50$ and $n = 12,000$ imply $R^2 \approx 0.001685$, slightly different from the printed 0.0016. Use the supplied R^2 and the corresponding 0.16% for Questions 19–21; the interpretation is unchanged.

Labels: The complete regression output refers to Dataset 5, although its introductory sentence says Dataset 6. The exam skips the Dataset 7 label. The opening p-value question is unnumbered; there are 26 numbered questions.

General standard: Reward scientifically correct interpretation, not the presence of the word "significant." Accept justified test choices, equivalent wording, appropriate caveats, and reasonable rounding.