BCB744 Biostatistics — Theory Test (Version 1)

Total: 100 marks | Time: 90 minutes

A. J. Smit
University of the Western Cape

2026-01-01

> ! Important
>
> **Instructions**
>
> - This paper has **three parts**: Part A (General Theory, 50 marks), Part B (Experiment Design and Hypothesis Formulation, 25 marks), and Part C (Statistical Output Interpretation, 25 marks).
> - Answer **all** questions.
> - Write clearly and in complete sentences where prose is required.
> - Mark allocations are shown next to each question in **(/ marks)** notation.
> - Statistical notation: use H_0 for the null hypothesis and H_A for the alternative hypothesis.

Part A: General Theory (50 marks)

Question 1 — The Scientific Method (/6)

a. Explain the difference between a null hypothesis and an alternative hypothesis. **(/ 2)**
b. Why is it important to formulate hypotheses *before* collecting data? What statistical problem arises when hypotheses are adjusted after seeing the data? **(/ 2)**
c. What is a confounding variable? Provide one example from biology and explain how you would control for it in an experiment. **(/ 2)**

Question 2 — Descriptive Statistics and Visualisation (/5)

a. When is the median a more appropriate measure of central tendency than the mean? Give a biological example where you would choose the median. **(/ 2)**
b. A researcher has continuous measurements of body mass (in grams) from three species of lizard. Suggest the most informative plot type for comparing the distributions across species, and explain in one sentence why it is better than a bar chart with error bars. **(/ 3)**

Question 3 — Probability Distributions and the Central Limit Theorem (/8)

a. List **three** characteristics of the normal distribution. **(/ 3)**
b. A researcher is counting the number of bird nesting attempts per territory per season. Under what conditions would you expect these counts to follow a Poisson distribution? What key assumption must hold? **(/ 3)**
c. State the Central Limit Theorem and explain why it is important for applying parametric hypothesis tests to biological data that are not perfectly normally distributed. **(/ 2)**

Question 4 — Statistical Inference and Error (/7)

a. Define a p-value in plain language (without using the word "probability" in a circular way). **(/ 2)**
b. Distinguish between a Type I error and a Type II error. Which one does the significance level α directly control? **(/ 3)**
c. A researcher increases the sample size of their experiment from $n = 20$ to $n = 80$. What effect does this have on statistical power, and why? **(/ 2)**

Question 5 — Assumptions and Transformations (/6)

a. List the **three** main assumptions that must hold before applying a parametric test (such as a t-test or ANOVA). For each, name one diagnostic method. **(/ 6)**

Question 6 — Correlation and Association (/8)

a. Explain the conceptual difference between Pearson's r and Spearman's ρ (rho). Under what circumstances would you choose Spearman over Pearson? **(/ 4)**
b. A marine biologist finds a strong positive correlation ($r = 0.91$) between mean sea surface temperature (SST) and the frequency of coral bleaching events over a 20-year dataset. A colleague claims this proves that warming SST *causes* bleaching. Provide **two** alternative explanations for this correlation that do not require direct causation, and briefly explain why the biologist's claim is premature. **(/ 4)**

Question 7 — Simple Linear Regression (/10)

a. A regression of seagrass shoot density (shoots m^{-2}) on water clarity (Secchi depth, m) yields: $\hat{y} = 14.3 + 8.7x$. Interpret both the intercept and the slope in biological terms. **(/ 2)**
b. The model returns $R^2 = 0.71$. What does this value mean? **(/ 2)**
c. Describe **three** diagnostic plots or tests you would use to verify that the assumptions of the linear regression model are met. For each, state what assumption it checks and what a violation would look like. **(/ 3)**
d. Explain the difference between a **confidence interval** and a **prediction interval** for a regression model. Which is wider, and why? **(/ 3)**

Part B: Experiment Design and Hypothesis Formulation (25 marks)

Question 8 — Seabird Egg Mass Across Island Populations (/12)

A researcher is studying egg mass (g) of a colonial seabird across four island populations in the sub-Antarctic. The first six rows of the dataset are shown below:

```
  island  egg_mass_g
1      A       82.4
2      A       79.1
3      A       84.3
4      B       91.7
5      B       88.2
6      B       94.5
```

The dataset contains 72 records: 18 eggs measured from each of the four islands (A, B, C, D). The researcher wishes to determine **whether mean egg mass differs significantly among the four island populations**.

a. State the formal null and alternative hypotheses for this analysis (use appropriate notation). **(/ 3)**
b. Identify the most appropriate statistical test for this research question. **(/ 2)**
c. Provide **three** specific reasons why you selected this test, with reference to the nature of the predictor and response variables, the number of groups, and the assumptions required. **(/ 6)**
d. If the main test returns a significant result, what additional procedure would you perform, and why? **(/ 1)**

Question 9 — Intertidal Algal Cover and Wave Exposure (/13)

An ecologist samples 50 intertidal rocky-shore plots across a gradient of wave exposure. For each plot, wave exposure is measured on a continuous index (0 = fully sheltered; 100 = fully exposed) and percentage cover of the dominant alga *Ectocarpus* sp. is estimated. The first six rows of the dataset are:

```
  plot_id  exposure_index  cover_pct
1       1             4.2       71.3
2       2             9.8       68.9
3       3            18.5       59.4
4       4            31.2       47.1
5       5            47.6       38.2
6       6            58.1       29.5
```

The researcher's aim is: *"To determine whether there is a significant relationship between wave exposure and the cover of* Ectocarpus *sp. in the intertidal zone."*

a. State formal null and alternative hypotheses appropriate for this research aim. **(/ 3)**
b. Identify the most appropriate statistical test and state **two** reasons for your choice, with reference to the nature of both variables. **(/ 4)**
c. What **two** assumption checks would you perform *before* fitting the model, and what would you do if each assumption were violated? **(/ 4)**
d. Based on the data preview, describe what you expect the scatter plot to look like (direction, strength, linearity), and why this is consistent with the biological interpretation. **(/ 2)**

Part C: Statistical Output Interpretation (25 marks)

Question 10 — Wilcoxon Rank-Sum Test Output (/12)

A researcher compares the percentage cover of lichen on north-facing versus south-facing granite outcrops at 20 sites per aspect. The data failed the Shapiro-Wilk normality test. The following output was produced:

```
    Wilcoxon rank-sum exact test

data:  cover_pct by aspect
W = 142, p-value = 0.0234
alternative hypothesis: true location shift is not equal to 0

95 percent confidence interval:
  1.450  12.351
sample estimates:
difference in location
                6.912
```

a. State the null hypothesis that this test evaluated. **(/ 2)**
b. What does the test statistic $W = 142$ represent conceptually? **(/ 2)**
c. Interpret the p-value = 0.0234 at $\alpha = 0.05$. What conclusion do you draw? **(/ 3)**
d. What does the 95% confidence interval (1.450, 12.351) tell you, in plain language? **(/ 3)**
e. Why was the Wilcoxon rank-sum test used instead of an independent-samples t-test? **(/ 2)**

Question 11 — Two-Way ANOVA Output (/13)

An experiment tests the effects of temperature (three levels: 15°C, 20°C, 25°C) and nutrient addition (two levels: ambient, enriched) on the growth rate (cm day^{-1}) of a marine macroalga. Ten replicates per treatment combination were used. The ANOVA table is:

```
                    Df Sum Sq Mean Sq F value   Pr(>F)
temperature          2  845.3   422.7  18.432  < 0.001 ***
nutrient             1  312.1   312.1  13.610   0.0004 ***
```

```
temperature:nutrient    2   89.4    44.7   1.950   0.1503
Residuals              54 1238.7    22.9

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

a. Identify the two predictor variables in this experiment. **(/ 2)**
b. Which main effects are statistically significant? Which interaction term is significant? **(/ 3)**
c. Interpret the non-significant interaction term biologically. What does this mean for how temperature and nutrient effects operate on algal growth? **(/ 3)**
d. What is the residual mean square (22.9) measuring in this experiment? **(/ 2)**
e. The *F*-value for temperature is 18.432. Show how it was calculated from the ANOVA table, and explain what it measures. **(/ 3)**

End of Version 1

Bibliography