BCB744: Biostatistics R Test

Smit, A. J.

2025-04-25

About the Test

The Biostatistics Test will start at 8:30 on 25 April, 2025 and you have until 11:30 to complete it. This is the Theory Test, which must be conducted on campus. The theory component contributes 30% of the final assessment marks.

Assessment Policy

The marks indicated for each section reflect the relative weight (and hence depth expected in your response) rather than a rigid check-list of individual points. Your answers should demonstrate a comprehensive understanding of the concepts and techniques required. Higher marks will be awarded for narratives that demonstrate not only conceptual and theoretical correctness but also insightful discussion and clear communication of insights or findings. We are assessing your ability to think systematically through complex inquiries, make appropriate theoretical and methodological choices, and present feedback in a coherent narrative that reveals deep understanding.

Please refer to the Assessment Policy for more information on the test format and rules.

Theory Test

This is the closed book assessment.

Below is a set of questions to answer. You must answer all questions in the allocated time of 3-hr. Please write your answers in a neatly formatted Word document and submit it to the iKamva platform.

Clearly indicate the question number and provide detailed explanations for your answers. Use Word's headings and subheadings facility to structure your document logically.

Naming convention: `Biostatistics_Theory_Test_YourSurname.docx`

Question 1

Imagine you are presented with the following five research scenarios (see below). In each case, your task is to decide which statistical method would be most appropriate and to justify your reasoning.

For each of the five scenarios below:

a. Identify the appropriate statistical method.
b. Explain why this method is more suitable than the others listed.

c. Clearly identify the dependent and independent variables (where applicable), and describe their type (categorical, continuous, etc.).
d. Describe what the method would allow you to infer, and what its limitations might be in the given context.

Scenarios:

1. A researcher wants to compare average leaf nitrogen content between two plant species growing in the same habitat.
2. An ecologist is interested in whether water temperature predicts fish body size across multiple river sites.
3. A conservation biologist is comparing average bird abundance across five habitat types, while also accounting for altitude which is known to influence bird detection rates.
4. A physiologist wants to explore whether heart rate and body temperature are linearly associated in a sample of animals under heat stress conditions.
5. A botanist tests whether fertiliser type (3 levels: organic, inorganic, control) affects plant height, but only has access to a small sample from each group.

[20 marks]

Answer

Scenario 1:

a. Independent (two-sample) *t*-test (or Mann-Whitney U test if data are not normally distributed).
b. The *t*-test is appropriate for comparing means between two independent groups (species). The Mann-Whitney U test is a non-parametric alternative that does not assume normality.
c. Dependent variable: leaf nitrogen content (continuous); independent variable: plant species (categorical).
d. The *t*-test allows for inference about differences in means, but is sensitive to normality and equal variance assumptions. The Mann-Whitney U test is less sensitive to these assumptions but does not provide mean differences (differences based on ranks).

Scenario 2:

a. Linear regression analysis.
b. Linear regression is suitable for assessing the relationship between a continuous dependent variable (fish body size) and a continuous independent variable (water temperature).
c. Dependent variable: fish body size (continuous); independent variable: water temperature (continuous).
d. Linear regression allows for inference about the strength and direction of the relationship, but assumes linearity and homoscedasticity. It may not capture non-linear relationships or interactions.

Scenario 3:

a. Analysis of covariance (ANCOVA).
b. ANCOVA is appropriate for comparing means across multiple groups (habitat types) while controlling for a covariate (altitude).
c. Dependent variable: bird abundance (continuous); independent variable: habitat type (categorical); covariate: altitude (continuous).
d. ANCOVA allows for inference about group differences while accounting for the influence of altitude, but assumes homogeneity of regression slopes and normality of residuals.

Scenario 4:

a. Linear regression analysis.
b. Linear regression is suitable for exploring the relationship (often causal) between two continuous variables (heart rate and body temperature).
c. Dependent variable: heart rate (continuous); independent variable: body temperature (continuous).
d. Linear regression allows for inference about the strength and direction of the relationship, but assumes linearity and homoscedasticity. It may not capture non-linear relationships or interactions.

Scenario 5:

a. One-way ANOVA (or Kruskal-Wallis test if data are not normally distributed).
b. One-way ANOVA is appropriate for comparing means across three or more independent groups (fertiliser types). The Kruskal-Wallis test is a non-parametric alternative that does not assume normality.
c. Dependent variable: plant height (continuous); independent variable: fertiliser type (categorical).
d. One-way ANOVA allows for inference about differences in means across groups, but assumes normality and homogeneity of variances. The Kruskal-Wallis test is less sensitive to these assumptions but does not provide mean differences (differences based on ranks).

Assessment Criterion	**Descriptor**	**Marks**
1. Conceptual grasp of epistemology	Shows understanding of epistemology as a theory of knowledge: how we know, not just what we know. Distinguishes epistemic claims from metaphysical or moral ones. *0–1:* Misunderstands or omits the concept. *2–3:* Partial understanding; conflates with method or belief. *4–5:* Clearly articulates epistemology in context; shows reflective engagement.	0–5
2. Explanation of scientific epistemic structure	Demonstrates understanding of how the scientific method justifies knowledge: e.g., empirical observation, theory-laden verification, replication, falsifiability, and provisionality. *0–1:* Provides no or incorrect explanation. *2–3:* References features like evidence or experiments, but lacks structure. *4–5:* Offers a coherent account of how science generates and revises claims.	0–5
3. Contrast with faith-based systems	Identifies how religious or mystical traditions legitimise knowledge through authority, revelation, or inner conviction. Avoids caricature or simplification. *0–1:* Simplistic binary (e.g., "science = truth, religion = belief"). *2–3:* Describes basic contrast but misses nuance. *4–5:* Analyses contrasts in justification, verification, and correction.	0–5
4. Integration of material from assigned reading	Effectively engages with relevant material from the assigned chapter (e.g., Galileo's telescope, instrument-based trust, constructivist critiques). *0:* No reference to or engagement with the text. *1:* Superficial mention without integration. *2–3:* Clear synthesis of reading into argument.	0–3

5. Quality of argumentation and writing	Clarity, structure, and originality of response. Logical progression, precise language, and effective use of examples. *0:* Incoherent or poorly expressed. *1:* Reasonably clear with occasional lapses. *2:* Fluent, well-organised, and compelling.	0–2
	Total	**20**

Question 2

Science does not rely on certainty but on scepticism and structured doubt. Its premise is not the claim to final truth; rather, it has the capacity to generate reliable, revisable knowledge through empirical observation, theoretical coherence, and methodological transparency.

In contrast, faith-based systems appeal to revelation, authority, or moral intuition – forms of conviction that do not invite or value independent verification. Yet both systems organise trust. What, then, distinguishes scientific knowledge from belief? What makes the scientific method a unique epistemological endeavour?

Question: What is the basis of knowledge in the scientific method, and how does this differ from the basis of knowledge in faith-based systems such as religion or mysticism? In your answer, consider the roles of observation, verification, theoretical coherence, and error correction in scientific reasoning, and contrast these with how knowledge is "made real" in non-empirical approaches.

[15 marks]

Answer

Assessment Criterion	Descriptor	Marks
1. Epistemological basis of scientific method	Identifies how science generates and legitimises knowledge through observation, verification, coherence, error correction, and structured doubt. *0–1:* Fails to explain or conflates epistemology with method or belief. *2–3:* Some understanding of empirical structure, but lacks clarity or depth. *4:* Coherent account of science's epistemological architecture.	0–4
2. Contrast with faith-based epistemologies	Explains how belief systems such as religion or mysticism ground knowledge in non-empirical sources (revelation, authority, moral intuition). *0–1:* Simplistic or dismissive contrast. *2:* Recognises distinction but lacks detail or nuance. *3:* Articulates key contrasts in verification, justification, and trust.	0–3
3. Use of key terms and concepts	Employs terms such as observation, verification, coherence, falsifiability, and "made real" in epistemically meaningful ways. *0–1:* Little or no use of relevant concepts. *2:* Some terminology used but inconsistently or unclearly. *3:* Accurate and conceptually integrated use of language.	0–3
4. Comparative insight and originality	Shows insight into how both systems organise trust and distinguish belief from knowledge. Avoids binary clichés. *0–1:* Uncritical or overly oppositional. *2:* Reasonable contrast, but surface-level. *3:* Offers reflective or original comparison of epistemic norms.	0–3

5. Coherence and written expression	Organised, precise, and cogent writing; ideas flow logically. *0:* Poorly expressed or incoherent. *1:* Understandable, but uneven. *2:* Clear, structured, and engaging.	0–2
	Total	**15**

Question 3

Throughout history, the development of statistical reasoning has been shaped not just by mathematical discoveries, but by synergies across intellectual traditions, technological innovation, and societal imperatives. From ancient record-keeping and proto-quantification, through the epistemic insights of the Renaissance and Enlightenment, to the formalisation of probabilistic thinking, statistics has evolved alongside shifting ideas about what it means to *know*, to *measure*, and to *infer*.

Question: How have historical interactions between these forces – ideas, instruments, and institutions – shaped the philosophy underpinning statistical practice as we know it today? In your response, identify and critically examine what you consider, with justification, to be five major conceptual or methodological turning points. These may include developments in logical reasoning, technological breakthroughs that extended observational capacity, institutional needs for demographic governance, or shifts in philosophical approaches to uncertainty and knowledge.

Your analysis should not simply recount historical facts, but provide a reasoned argument about how each moment contributed to the emergence of statistics as a knowledge framework – that is, not just a set of techniques, but a way of thinking about the world.

[20 marks]

Answer

Assessment Criterion	Descriptor	Marks
1. Identification and justification of five turning points	Selects five relevant developments (conceptual, methodological, technological, institutional) and justifies their importance in shaping statistical thought. *0–2:* Incomplete or poorly justified selection. *3–4:* Reasonable choices, with limited justification. *5:* Clear, well-motivated and historically grounded selection.	0–5
2. Explanation of interactions among ideas, tools, and institutions	Demonstrates how intellectual, technological, and societal forces interacted to shape statistical philosophy. *0–1:* Fragmented account with little synthesis. *2–3:* Recognises key interactions but lacks depth or integration. *4–5:* Coherent analysis of mutual reinforcement and historical context.	0–5
3. Engagement with epistemological concepts	Articulates how statistical reasoning relates to ideas of uncertainty, inference, observation, and measurement. *0–1:* Superficial or absent treatment. *2–3:* Some conceptual reflection, but underdeveloped. *4–5:* Strong engagement with epistemic foundations.	0–5
4. Use of assigned reading and historical material	Integrates material from the chapter (e.g., Galileo, the printing press, van Leeuwenhoek, Laplace, etc.) to support the argument. *0–1:* Little or no reference to the reading. *2:* Uses examples but with minimal integration. *3:* Demonstrates meaningful synthesis with the source material.	0–3

5. Coherence, structure, and originality	Writing is well-organised and shows independent thought. Argument flows logically, with appropriate variation in style and pace. *0:* Disorganised or difficult to follow. *1:* Generally coherent, but uneven. *2:* Clear and competent. *3:* Persuasive, well-paced, and conceptually engaging.	0–3
	Total	**20**

Question 4

Statistical reasoning begins with our wish to learn about something large and often inaccessible by examining something smaller and manageable. The credibility of this approach – from observed data to broader inference – depends on how we conceptualise and structure the relationship between what we observe and what we want to know.

This question asks that you examine the important terms and principles that make this act of inference possible.

Question: What do statisticians mean by "population" and "sample"? Define each term clearly, and explain the distinction between them. How are they related in practice, and how does the method of sampling affect the validity of estimates for population parameters such as the mean and dispersion? Support your discussion with examples where appropriate.

[10 marks]

Answer

Assessment Criterion	Descriptor	Marks
1. Definition and distinction: population vs. sample	Provides clear, accurate definitions. Demonstrates understanding of how a sample is conceptually and inferentially linked to a population. *0:* Definitions absent or incorrect. *1:* Basic or vague explanation. *2:* Clear, accurate, and well-articulated definitions and distinctions.	0–2
2. Relationship in practice	Explains how samples are used to draw conclusions about populations; identifies the rationale for using samples. *0:* No explanation or incorrect claim. *1:* Partial understanding. *2:* Correct and practically contextualised explanation.	0–2
3. Role of sampling method	Identifies how sampling strategies (random, biased, etc.) influence the reliability of estimates like the mean and dispersion. *0:* No discussion of sampling method. *1:* Mentions method but lacks detail. *2:* Analytically explains how sampling quality affects inferential accuracy.	0–2
4. Impact on estimates of population parameters	Connects sampling quality to estimates of central tendency and variation. May address bias, variability, or representativeness. *0:* No reference to estimation. *1:* Simplistic account (e.g., just states "affects accuracy"). *2:* Well-reasoned explanation with statistical relevance.	0–2

5. Use of relevant examples and clarity of expression	Supports discussion with apt examples; communicates ideas clearly and logically. *0:* Unclear or no examples. *1:* Example provided but not integrated. *2:* Well-chosen example(s) that enhance explanation.	0–2
	Total	**10**

Question 5

Words shape our thoughts, and nowhere is this more consequential than in science, where terminological precision goes hand-in-hand with conceptual clarity. Statistical terms like "random" or "stochastic" carry specific meanings in the context of probabilistic logic and mathematical formalism. Yet in everyday language, such terms are often misused. They are flattened into colloquialisms that only hint at their true meaning. This insidious slippage is more than semantic; it has consequences for how we value knowledge.

Why does it matter if "random" is used imprecisely? How do scientific concepts become confused, or even trivialised, when technical language is absorbed into everyday language without regard for its analytic structure?

Question: Discuss the scientific meaning of "random" and contrast it with its colloquial usage. Why is this distinction important for statistical reasoning, and how can imprecise language lead to conceptual misunderstandings? In your answer, consider how terms like "haphazard" and "unpredictable" differ from "random," and evaluate the knowledge implications of using such words loosely in scientific or public discourse.

[10 marks]

Answer

Assessment Criterion	Descriptor	Marks
1. Definition of scientific and colloquial meanings of "random"	Provides a clear and accurate definition of "random" in statistical reasoning and contrasts it meaningfully with everyday usage. *0:* Incorrect or missing definitions. *1:* Partial or imprecise contrast. *2:* Accurate definitions and well-articulated contrast.	0–2
2. Explanation of significance in statistical reasoning	Explains why conceptual precision around "randomness" matters for designing, interpreting, or trusting statistical inference. *0:* No justification or misunderstanding of significance. *1:* Basic relevance noted but not developed. *2:* Shows clear understanding of why terminological precision matters.	0–2
3. Discussion of related terms and conceptual confusion	Evaluates how terms like "haphazard" or "unpredictable" differ from "random," and discusses implications of terminological slippage. *0:* No mention of related terms. *1:* Terms mentioned but distinction not clearly drawn. *2:* Analytically distinguishes and explores conceptual confusion.	0–2

4. Evaluation of consequences for knowledge or discourse	Assesses how loose language affects scientific literacy or distorts public understanding. *0:* No evaluation of broader consequences. *1:* Mentions issue but lacks depth. *2:* Engages thoughtfully with implications for knowledge/policy/discourse.	0–2
5. Clarity, expression, and structure	Writing is coherent, conceptually organised, and shows linguistic control. *0:* Disorganised or opaque. *1:* Understandable but uneven. *2:* Clear, persuasive, and well-structured.	0–2
	Total	**10**

Question 6

Your task is to design a hypothetical study that could lead to a statistical analysis using one of the following methods:

- One-way ANOVA
- Simple linear regression
- Pearson or Spearman correlation

Your study may involve field sampling, a laboratory experiment, or observational data – what matters is that your design aligns meaningfully with the statistical method you choose.

In your answer, do the following:

1. Describe your hypothetical experiment or sampling campaign.
    - Outline what you are investigating, how data will be collected, and what your variables are. Be clear about their measurement scale (categorical, continuous) and expected behaviour.
    - Present this as a formally written Methods section suitable for a peer-review publication.
2. Justify the statistical method you have chosen.
    - Explain why your design is appropriate for ANOVA, regression, or correlation.
3. Formally state the null and alternative hypotheses as they would be tested in the chosen analysis.
4. Show a portion of the pseudo-data as one would see using the `head()` or `tail()` functions in R.
    - This should be a small, representative sample of the data you would collect.
5. Describe the sequence of analytical steps you would take – from raw data to final conclusion.
    - Include any relevant assumptions, diagnostic checks, or transformations that may be required before interpreting the results.
6. Write a hypothetical Results section that summarises the findings of your analysis.
    - This should include a brief interpretation of the statistical output, including relevant pseudo-tables or pseudo-figures.

Your answer should reflect an understanding of the logic and structure of statistical inference, from design to decision. You are welcome to use R and RStudio to generate any data, tables, and graphs, should you wish.

[25 marks]

Answer

Assessment Criterion	Descriptor	Marks

1. Experimental or sampling design (Methods section)	Presents a hypothetical study with clear variables, measurement types (categorical/continuous), and logical data collection approach. Framed in the style of a peer-reviewed Methods section. *0–2:* Vague, underdeveloped, or incoherent. *3–4:* Adequate design, partially formalised. *5:* Clear, plausible, and professionally structured.	0–5
2. Justification of chosen statistical method	Explains why the method (ANOVA, regression, correlation) is appropriate based on the variables and study question. *0–1:* Method chosen without justification. *2–3:* Method mostly appropriate, with some justification. *4:* Method fully justified and aligned with design.	0–4
3. Hypothesis formulation	States null and alternative hypotheses as they would appear in formal statistical testing. Correctly aligned with method and data structure. *0–1:* Incorrect or missing. *2:* Present but informal or poorly structured. *3:* Formally correct and clearly expressed.	0–3
4. Representative data sample	Includes a small, clearly formatted table of (pseudo-)data to illustrate variables. May simulate 'head()' or 'tail()' output. *0–1:* Absent or irrelevant data. *2:* Present but unstructured or unclear. *3:* Representative and appropriately formatted.	0–3
5. Analytical workflow (from raw data to inference)	Describes logical steps: assumptions, transformations, model diagnostics, and inferential strategy. *0–1:* Minimal or confused. *2–3:* Partial sequence, some omissions. *4:* Coherent, technically sound workflow.	0–4
6. Interpretation and Results summary	Provides a hypothetical Results section interpreting the (pseudo-)statistical outcome, with mention of output tables/figures. *0–1:* No interpretation or incoherent. *2–3:* Interprets outcome, but superficially. *4–5:* Thoughtful, succinct summary with clear output reference.	0–5
	Total	**25**

TOTAL MARKS: 100

– THE END –