A Level Statistics
Practice Test 6: Data Collection

Instructions:
Answer all questions. Show your working clearly.
Calculators may be used unless stated otherwise.
Draw diagrams where appropriate to illustrate your solutions.
Time allowed: 3 hours

Section A: Modern Data Collection Methods [25 marks]

1. **[8 marks]** Define and explain contemporary data collection approaches:

   (a) Define big data and explain how it differs from traditional data collection.

   (b) Explain web scraping and its applications in modern research.

   (c) Describe sensor data collection and provide examples.

   (d) Define administrative data and explain its advantages for research.

2. **[10 marks]** Compare traditional and modern data collection methods for these scenarios:

   (a) Studying consumer behavior patterns

   (b) Monitoring traffic flow in a city

   (c) Understanding social media influence on political opinions

   (d) Tracking health outcomes in a population

   (e) Measuring customer satisfaction

   (f) Analyzing educational achievement

   (g) Studying environmental changes

   (h) Investigating workplace productivity

   (i) Monitoring financial market trends

   (j) Researching travel and tourism patterns

3. **[7 marks]** A retail company uses multiple data sources:

   - Point-of-sale transaction data (automated)
   - Customer loyalty card records (administrative)
   - Website browsing analytics (digital tracking)
   - Customer service call logs (operational data)

- Social media mentions and reviews (web scraping)

(a) Classify each data source as structured or unstructured.

(b) Identify potential advantages of each data source.

(c) Describe challenges in combining these different data types.

Section B: Data Quality and Validation [30 marks]

4. **[12 marks]** Explain data quality assessment and validation techniques:

(a) Define data completeness, accuracy, consistency, and timeliness.

(b) Describe methods for detecting missing data patterns.

(c) Explain cross-validation techniques for data verification.

(d) Describe outlier detection methods and their appropriate use.

5. **[18 marks]** A health monitoring app collects daily step count data from 1000 users over 30 days. Initial analysis reveals:
- 15- Some users report impossibly high step counts (¿100,000 steps/day) - Weekend step counts are systematically 30- 50 users have zero recorded steps for entire weeks - Step counts jump dramatically when users switch from phone to fitness tracker

(a) Identify and classify each data quality issue.

(b) Calculate the overall data completeness rate.

(c) Suggest appropriate methods to handle missing data.

(d) Propose criteria for identifying and handling outliers.

(e) Design validation rules for step count data.

(f) Recommend methods to ensure data consistency across devices.

(g) Suggest ways to improve data collection procedures.

(h) Calculate what percentage of users have complete data records.

(i) Propose a data quality score system for this dataset.

Section C: Comprehensive Data Analysis Project [35 marks]

6. **[15 marks]** Design a comprehensive study to investigate factors affecting student academic performance:

(a) Define clear research objectives and hypotheses.

(b) Identify the target population and sampling frame.

(c) Design an appropriate sampling strategy with justification.

(d) List key variables to collect and classify each by type.

(e) Choose suitable data collection methods for each variable type.

7. **[20 marks]** Analyze this multi-source dataset about online learning effectiveness:

Data Source 1 - Survey Responses (n=500): - Student satisfaction ratings (1-5 scale) - Weekly study hours (self-reported) - Technology access level (Low/Medium/High)

Data Source 2 - Learning Platform Analytics: - Time spent on platform (minutes/week) - Assignment submission rates (- Video lecture completion rates (

Data Source 3 - Academic Records: - Final course grades (0-100- Previous academic performance (GPA) - Course difficulty level (1-5 scale)

Key findings from initial analysis: - Strong positive correlation (r=0.78) between platform time and grades - Self-reported study hours 40- High technology access students score 15- 25- Assignment submission rates vary from 45

(a) Identify potential data quality issues in this multi-source dataset.

(b) Explain the discrepancy between self-reported and actual study time.

(c) Suggest methods to validate and triangulate findings across data sources.

(d) Design appropriate visualizations for each type of finding.

(e) Propose additional data that would strengthen the analysis.

(f) Identify potential confounding variables that should be considered.

(g) Recommend actions based on the key findings.

(h) Assess the reliability and validity of conclusions from this study.

Answer Space

Use this space for your working and answers.

Formulae and Key Concepts

Data Quality Dimensions:

Completeness = $\frac{\text{Non-null values}}{\text{Total values}} \times 100\%$

Accuracy = $\frac{\text{Correct values}}{\text{Total values}} \times 100\%$

Consistency = Degree of uniformity across data sources

Timeliness = Freshness and relevance of data

Missing Data Patterns:

MCAR: Missing Completely At Random

MAR: Missing At Random (depends on observed data)

MNAR: Missing Not At Random (depends on unobserved data)

Outlier Detection Methods:

Statistical: Z-score, IQR method, Grubbs' test
Visual: Box plots, scatter plots, histograms
Machine learning: Isolation forests, clustering
Domain-specific: Business rules, expert knowledge

Data Validation Techniques:
Range checks: Values within expected bounds
Format checks: Correct data types and patterns
Consistency checks: Logical relationships maintained
Completeness checks: Required fields populated
Cross-reference checks: External data verification

Modern Data Characteristics:
Volume: Scale of data (terabytes, petabytes)
Velocity: Speed of data generation and processing
Variety: Different types and formats of data
Veracity: Quality and reliability of data
Value: Potential insights and benefits

Data Source Types:
Structured: Organized in tables/databases
Semi-structured: JSON, XML formats
Unstructured: Text, images, videos
Real-time: Continuous data streams
Batch: Periodic data collection

Research Design Elements:
Population: Target group of interest
Sample: Subset selected for study
Variables: Characteristics being measured
Methodology: Data collection procedures
Analysis plan: Statistical techniques to be used

Data Integration Challenges:
Schema matching: Aligning different data structures
Entity resolution: Identifying same entities across sources
Data fusion: Combining conflicting information
Temporal alignment: Synchronizing time-series data
Quality harmonization: Standardizing quality levels

END OF TEST

Total marks: 90

For more resources and practice materials, visit:
stepupmaths.co.uk