A Level Statistics
Practice Test 6: Measures of Location and Spread

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: Bootstrap and Resampling Methods [25 marks]

1. **[12 marks]** Define and apply bootstrap methodology to statistical estimation:

(a) Define the bootstrap method and explain its purpose in statistical inference.

(b) Explain how bootstrap samples are created from an original dataset.

(c) Describe the relationship between bootstrap sampling and the concept of sampling with replacement.

(d) For the dataset [12, 15, 18, 22, 25], list five possible bootstrap samples of size 5.

(e) Calculate the mean of each bootstrap sample from part (d).

(f) Explain how the distribution of bootstrap means estimates the sampling distribution of the original mean.

2. **[8 marks]** Apply bootstrap methods to estimate confidence intervals:

(a) Explain the bootstrap percentile method for constructing confidence intervals.

(b) Given 1000 bootstrap means with 2.5th percentile = 14.2 and 97.5th percentile = 18.6, construct a 95

(c) Compare bootstrap confidence intervals with traditional t-distribution methods.

(d) Describe when bootstrap methods are particularly advantageous over traditional approaches.

3. **[5 marks]** Analyze bootstrap applications for non-standard statistics:

(a) Explain how bootstrap can estimate the standard error of the median.

(b) Describe bootstrap applications for estimating confidence intervals of correlation coefficients.

(c) Discuss the limitations of bootstrap methods and when they might not be appropriate.

Section B: Bayesian Statistics and Prior-Posterior Analysis [30 marks]

4. **[15 marks]** Define and apply Bayesian approaches to statistical estimation:

(a) Define prior distribution, likelihood, and posterior distribution in Bayesian statistics.

(b) Explain Bayes' theorem: $P(\theta|data) = \frac{P(data|\theta) \times P(\theta)}{P(data)}$

(c) Describe how prior beliefs are updated with new data to form posterior beliefs.

(d) A coin has unknown probability p of heads. Prior belief: p Uniform(0,1). After 10 flips showing 7 heads, describe how to update the posterior.

(e) Explain the concept of conjugate priors and their computational advantages.

(f) Compare Bayesian credible intervals with frequentist confidence intervals.

5. **[15 marks]** Apply Bayesian updating to a quality control scenario:

A factory produces electronic components with unknown defect rate . Historical data suggests follows a Beta(2,18) prior distribution (mean 0.1).

A new batch inspection reveals 3 defects in 50 components tested.

(a) Calculate the prior mean and variance for the defect rate.

(b) Apply Bayesian updating: the posterior is Beta(2+3, 18+50-3) = Beta(5,65).

(c) Calculate the posterior mean and variance.

(d) Compare how the new data has updated beliefs about the defect rate.

(e) Calculate the 95

(f) Determine the probability that the true defect rate exceeds 10

(g) Explain how this Bayesian approach differs from classical hypothesis testing.

(h) If 10 more components are tested with 1 defect, update the posterior again.

(i) Discuss how the precision of estimates changes with more data.

Section C: Machine Learning and Big Data Statistics [35 marks]

6. **[18 marks]** Analyze statistical measures in high-dimensional data contexts:

A machine learning model analyzes customer behavior using 500 features. Summary statistics for key features are:

Age: Mean=42.3, SD=12.8, Skewness=0.15, n=10,000 **Income:** Mean=£54,200, SD=£18,500, Skewness=1.24, n=10,000 **Purchase Frequency:** Mean=2.8/month, SD=1.6, Skewness=-0.31, n=10,000

(a) Calculate confidence intervals for the population means of each feature (use z-distribution for large n).

(b) Assess which features show significant departure from normality based on skewness.

(c) Calculate the coefficient of variation for each feature and interpret relative variability.

(d) Standardize each feature (z-score transformation) and explain the benefits for machine learning.

(e) Estimate the 95

(f) Identify potential outliers for each feature using the 3-sigma rule.

(g) Explain how robust statistics might be more appropriate for the Income feature.

(h) Calculate the effective sample size for detecting a 5

(i) Discuss challenges in interpreting traditional statistics with 500-dimensional data.

7. **[17 marks]** Apply advanced statistical concepts to streaming data analysis:
An online platform processes streaming data with the following characteristics: - 1 million data points per hour - Rolling 24-hour statistics updated every minute - Real-time outlier detection required - Memory constraints limit storage to recent 10,000 points

(a) Explain how to maintain running means and variances efficiently as new data arrives.

(b) Describe the exponentially weighted moving average (EWMA) method: $S_t = \alpha X_t + (1-\alpha)S_{t-1}$.

(c) Calculate EWMA values for =0.1 with initial data: 100, 105, 98, 112, 95, 108.

(d) Compare EWMA with simple moving averages for trend detection.

(e) Design a real-time outlier detection system using statistical process control.

(f) Calculate dynamic control limits that adapt to changing data patterns.

(g) Explain how to estimate percentiles from streaming data using reservoir sampling.

(h) Discuss the trade-offs between accuracy and computational efficiency in streaming statistics.

(i) Design a system to detect significant changes in the data distribution over time.

(j) Propose methods for handling concept drift in the underlying data patterns.

Answer Space

Use this space for your working and answers.

Formulae and Key Concepts

Bootstrap Methods:

Bootstrap sample: Random sample with replacement from original data
Bootstrap statistic: Statistic calculated from bootstrap sample
Bootstrap percentile CI: Use percentiles of bootstrap distribution
Standard error estimate: SD of bootstrap statistics

Bayesian Statistics:

Bayes' theorem: $P(\theta|data) = \frac{P(data|\theta) \times P(\theta)}{P(data)}$
Beta distribution: $Beta(\alpha, \beta)$ with mean $\frac{\alpha}{\alpha+\beta}$
Bayesian updating: $Beta(\alpha + successes, \beta + failures)$

Credible interval: Bayesian equivalent of confidence interval

Exponentially Weighted Moving Average:

$EWMA_t = \alpha X_t + (1-\alpha)EWMA_{t-1}$

where = smoothing parameter (0 ¡ 1)

Higher gives more weight to recent observations

Variance: $\sigma^2_{EWMA} = \frac{\alpha}{2-\alpha}\sigma^2$

Running Statistics:

Running mean: $\bar{x}_n = \bar{x}_{n-1} + \frac{x_n - \bar{x}_{n-1}}{n}$

Running variance: Welford's algorithm for numerical stability

$M_2 = M_2 + (x_n - \bar{x}_{n-1})(x_n - \bar{x}_n)$

$s^2 = \frac{M_2}{n-1}$

High-Dimensional Statistics:

Curse of dimensionality: Distance measures become less meaningful

Feature standardization: $z = \frac{x-\mu}{\sigma}$

Robust measures needed due to outlier prevalence

Multiple testing corrections required

Statistical Process Control:

Control limits: ± 3 (99.7CUSUM charts: Detect small persistent changes

EWMA charts: Detect gradual shifts

Run rules: Patterns indicating process changes

Streaming Data Concepts:

Reservoir sampling: Maintain random sample from stream

Sliding window: Statistics over recent time period

Concept drift: Changes in underlying data distribution

Online algorithms: Process data sequentially without storage

Confidence Intervals (Large Samples):

Mean: $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

Proportion: $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Effect Size and Power:

Cohen's d: $d = \frac{|\mu_1 - \mu_2|}{\sigma}$

Sample size: $n = \frac{2(z_{\alpha/2} + z_\beta)^2\sigma^2}{(\mu_1 - \mu_2)^2}$

Outlier Detection Methods:

Z-score: —z— ¿ 3 (or 2.5)

IQR method: Outside Q1-1.5×IQR or Q3+1.5×IQR

Modified z-score: Uses median and MAD

Isolation forest: Machine learning approach

END OF TEST

Total marks: 90

For more resources and practice materials, visit:

stepupmaths.co.uk