A Level Statistics
Practice Test 6: Hypothesis Testing

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: Survival Analysis and Specialized Tests [25 marks]

1. **[12 marks]** Define and explain survival analysis concepts:

(a) Define survival analysis and explain when it's used instead of traditional hypothesis tests.

(b) Explain censoring in survival data and different types of censoring.

(c) Define the survival function and hazard function.

(d) Explain the Kaplan-Meier estimator and its purpose.

(e) Describe the log-rank test for comparing survival curves.

(f) Explain Cox proportional hazards regression and its applications.

2. **[8 marks]** Explain specialized hypothesis tests for specific distributions:

(a) Describe tests for normality (Shapiro-Wilk, Anderson-Darling, Kolmogorov-Smirnov).

(b) Explain tests for exponential distribution and their applications.

(c) Describe tests for Poisson distribution assumption in count data.

(d) Explain how to test for uniformity in data distributions.

3. **[5 marks]** Analyze robust hypothesis testing methods:

(a) Define robust statistics and explain their advantages.

(b) Describe trimmed means and their use in hypothesis testing.

(c) Explain Welch's t-test for unequal variances.

Section B: Sequential and Adaptive Testing [30 marks]

4. **[15 marks]** Define sequential analysis and adaptive designs:

(a) Explain sequential hypothesis testing and its advantages over fixed-sample designs.

(b) Define stopping rules and their role in sequential testing.

(c) Describe the Sequential Probability Ratio Test (SPRT).

(d) Explain adaptive clinical trial designs and their ethical advantages.

(e) Define interim analysis and alpha spending functions.

(f) Describe group sequential methods and their applications.

5. **[15 marks]** A pharmaceutical company conducts a sequential clinical trial testing a new drug against a placebo. They plan interim analyses after every 50 patients:

Analysis 1 (n=50): Treatment success: 28/25, Placebo success: 15/25, p-value = 0.02 **Analysis 2 (n=100):** Treatment success: 58/50, Placebo success: 32/50, p-value = 0.008 **Analysis 3 (n=150):** Treatment success: 89/75, Placebo success: 51/75, p-value = 0.001

Efficacy boundary: = 0.05 with Bonferroni correction for 5 planned analyses Futility boundary: Power ¡ 20

(a) Calculate the adjusted significance level for each interim analysis using Bonferroni correction.

(b) Assess whether the trial should stop for efficacy at each analysis point.

(c) Calculate the effect sizes (difference in proportions) at each analysis.

(d) Explain the ethical considerations for continuing or stopping the trial.

(e) Calculate confidence intervals for the treatment effect at each stage.

(f) Discuss the advantages and disadvantages of this sequential approach.

(g) Compare this to a fixed-sample design with n=150 and explain the benefits.

(h) Calculate the number needed to treat (NNT) at each analysis.

(i) Make a recommendation about trial continuation or termination.

Section C: Machine Learning and Hypothesis Testing [35 marks]

6. **[18 marks]** Explain the intersection of machine learning and hypothesis testing:

(a) Describe the difference between prediction and inference in statistical modeling.

(b) Explain cross-validation and its role in model selection and testing.

(c) Define overfitting and explain how it affects hypothesis testing conclusions.

(d) Describe the bias-variance trade-off in the context of hypothesis testing.

(e) Explain regularization methods and their impact on statistical inference.

(f) Describe permutation importance and its use in feature significance testing.

(g) Explain the multiple testing problem in high-dimensional data analysis.

(h) Describe false discovery rate (FDR) control methods.

(i) Explain how to test for model significance in machine learning contexts.

7. **[17 marks]** A data science team analyzes customer behavior using 500 features to predict purchase behavior. Their analysis reveals:

Model Performance: - Training accuracy: 94- Validation accuracy: 78- Test accuracy: 76- Cross-validation accuracy: 77

Feature Significance (Top 10 features):

Feature	Importance Score	p-value	FDR-adjusted p
Age	0.15	0.001	0.025
Income	0.12	0.002	0.040
Purchase History	0.10	0.008	0.080
Website Time	0.09	0.015	0.094
Email Opens	0.08	0.025	0.125
Social Media	0.07	0.045	0.188
Location	0.06	0.060	0.214
Device Type	0.05	0.080	0.250
Seasonality	0.04	0.120	0.333
Referral Source	0.03	0.180	0.450

(a) Assess whether the model shows evidence of overfitting and explain your reasoning.

(b) Identify which features are statistically significant after FDR correction at $= 0.05$.

(c) Explain why multiple testing correction is crucial in this high-dimensional setting.

(d) Calculate the false discovery rate for the unadjusted p-values.

(e) Discuss the trade-off between prediction accuracy and statistical interpretability.

(f) Recommend additional validation methods to ensure robust inference.

(g) Explain how to test whether the overall model is significantly better than random guessing.

(h) Discuss the limitations of p-values in machine learning contexts.

(i) Propose methods to improve both prediction and inference reliability.

(j) Design a hypothesis testing framework for validating these customer insights.

Answer Space

Use this space for your working and answers.

Formulae and Key Concepts

Survival Analysis:

Survival function: $S(t) = P(T > t)$

Hazard function: $h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$

Kaplan-Meier estimator: $\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$

Log-rank test: Compares survival distributions between groups

Sequential Testing:

SPRT: Continue sampling until sufficient evidence for H or H

Alpha spending function: $\alpha(t)$ allocates Type I error across analyses

Bonferroni for k analyses: $\alpha_{adj} = \frac{\alpha}{k}$

O'Brien-Fleming boundary: More stringent early stopping

Multiple Testing Correction:

Bonferroni: $\alpha_{adj} = \frac{\alpha}{m}$ (conservative)

Benjamini-Hochberg (FDR): Control expected proportion of false discoveries

Holm-Bonferroni: Step-down procedure, less conservative than Bonferroni

FDR $= E\left[\frac{V}{R}\right]$ where V = false discoveries, R = total discoveries

Effect Size Measures:

Number Needed to Treat: $NNT = \frac{1}{|p_1 - p_2|}$

Relative Risk: $RR = \frac{p_1}{p_2}$

Odds Ratio: $OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$

Cohen's h for proportions: $h = 2(\arcsin\sqrt{p_1} - \arcsin\sqrt{p_2})$

Model Validation:

Cross-validation: Split data into k folds, train on k-1, test on 1

Bootstrap validation: Resample with replacement, assess performance

Permutation tests: Randomly permute labels to test model significance

AUC-ROC: Area under receiver operating characteristic curve

Robust Statistics:

Trimmed mean: Remove extreme values before calculating mean

Welch's t-test: $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

Degrees of freedom: Satterthwaite approximation for unequal variances

Normality Tests:

Shapiro-Wilk: Best for small samples (n ¡ 50)

Anderson-Darling: Good power for detecting departures from normality

Kolmogorov-Smirnov: Tests goodness of fit to any distribution

Q-Q plots: Visual assessment of normality

Machine Learning Metrics:

Bias-Variance decomposition: $MSE = Bias^2 + Variance + Noise$

Regularization parameter: controls model complexity

Feature importance: Permutation-based or model-specific measures

Overfitting indicators: Large gap between training and validation performance

High-Dimensional Testing:

Bonferroni becomes very conservative with many tests

FDR control more powerful than FWER control
Benjamini-Hochberg procedure: Order p-values, find largest k where $p_{(k)} \leq \frac{k}{m}\alpha$

END OF TEST

Total marks: 90

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