Week 2: Statistical Analysis

Heart Failure Survival Analysis

MDST Project

Winter 2026

Outline

1. Correlation Analysis
2. Statistical Tests
3. Multiple Testing Correction
4. Summary

Pearson Correlation

- Measures **linear relationship** between two variables
- Range: -1 to $+1$
  - $+1$: Perfect positive correlation
  - 0: No linear correlation
  - -1: Perfect negative correlation
- Formula: $r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$

Feature Correlations with DEATH_EVENT

Positive Correlations:

- serum_creatinine: 0.29
- age: 0.25

Negative Correlations:

- time: -0.53
- ejection_fraction: -0.27
- serum_sodium: -0.20

Key Insight:
Follow-up time has the strongest correlation with death event, but this is expected (patients who die have shorter follow-up).

T-Test

- Compares **means** between two groups
- Assumes: Normal distribution, equal variances
- Welch's t-test: Does not assume equal variances
- Null hypothesis: No difference between group means

When to use:

- Continuous data
- Comparing two groups (survived vs. died)

Mann-Whitney U Test

- **Non-parametric** alternative to t-test
- Does NOT assume normal distribution
- Compares **ranks** instead of means
- More robust to outliers

When to use:

- Data is not normally distributed
- Ordinal data or skewed distributions

Significant Features (p ¡ 0.05)

Feature	T-test p-value	Mann-Whitney p-value
time	2.3×10^{-22}	6.9×10^{-21}
ejection_fraction	9.6×10^{-6}	7.4×10^{-7}
age	4.7×10^{-5}	1.7×10^{-4}
serum_creatinine	6.4×10^{-5}	1.6×10^{-10}
serum_sodium	1.9×10^{-3}	2.9×10^{-4}

The Problem with Multiple Testing

- Testing 12 features at $\alpha = 0.05$
- Each test has 5% chance of false positive
- Expected false positives: $12 \times 0.05 = 0.6$
- **Family-wise error rate** increases with more tests

Solution: Adjust p-values to control false discovery rate (FDR)

Benjamini-Hochberg (FDR) Correction

- Controls the **False Discovery Rate**
- FDR = Expected proportion of false positives among rejected hypotheses
- Less conservative than Bonferroni correction
- Procedure:
  1. Rank p-values from smallest to largest
  2. Adjust: $p_{adj} = p \times \frac{n}{rank}$
  3. Compare adjusted p-values to α

After FDR Correction

Still significant (FDR ¡ 0.05):

- time
- ejection_fraction
- age
- serum_creatinine
- serum_sodium

Not significant after correction:

- high_blood_pressure, anaemia, diabetes, platelets, sex, smoking, creatinine_phosphokinase

Key Takeaways

1. **Correlation** measures linear relationships
2. **T-test** compares means (assumes normality)
3. **Mann-Whitney U** is non-parametric (no normality assumption)
4. **Multiple testing correction** is essential when testing many hypotheses
5. **5 features** are significantly different between survival groups

Next Week: Unsupervised Learning

- **PCA** (Principal Component Analysis)
    - Dimensionality reduction
    - Visualizing high-dimensional data
- **Clustering**
    - Finding natural groupings in data
    - K-means, hierarchical clustering

Exercises

1. Write a function to return features with significant p-values given a threshold
2. Implement Mann-Whitney U test for all features
3. Interpret the correlation heatmap

Resources:

- Scipy Stats: `https://docs.scipy.org/doc/scipy/reference/stats.html`
- Statsmodels: `https://www.statsmodels.org/`