Week 2: Statistical Analysis & Feature Importance

Heart Failure Survival Analysis

MDST Project

Winter 2026

Outline

1. Week 1 Recap
2. The Lady Tasting Tea
3. Correlation Analysis
4. Statistical Tests
5. Multiple Testing Correction
6. Feature Importance
7. Feature Variance & Multicollinearity
8. Summary

Quick Recap: Week 1 - Exploratory Data Analysis

Dataset Overview:

- 299 heart failure patients
- 13 features (12 predictors + target)
- Target: DEATH_EVENT (0/1)
- 68% survived, 32% died
- No missing values

Feature Types:

- **Continuous:** age, ejection_fraction, serum_creatinine, serum_sodium, time, platelets, CPK
- **Binary:** anaemia, diabetes, high_blood_pressure, sex, smoking

Week 1 Observations (Visual)

What we saw in the plots:

Differences between groups:

- Died: **Higher** serum_creatinine
- Died: **Lower** ejection_fraction
- Died: **Shorter** follow-up time
- Died: **Older** patients

No obvious differences:

- Diabetes (similar in both)
- Sex (similar in both)
- Smoking (similar in both)

This Week's Goal: Use **statistical tests** to confirm these observations with numbers!

From Visualization to Statistics

Week 1: EDA	**Week 2: Statistics**
"The boxplots look different"	"The difference is statistically significant ($p < 0.05$)"
"This feature seems important"	"Random Forest ranks it #1"
"These features look correlated"	"VIF = 1.3, no multicollinearity"

Key Insight: Visualization suggests, statistics confirms!

A Summer Afternoon in Cambridge, 1920s

"I can tell whether the milk was poured first, or the tea."

The Scene:
A group of scientists and academics gathered for afternoon tea at Cambridge University. Among them was Dr. Muriel Bristol, an algologist (scientist who studies algae).

When offered a cup of tea, she politely declined – insisting that she preferred her tea prepared with **milk poured into the cup first**, before the tea.

"I can taste the difference," she claimed.

The scientists laughed. **Surely that's impossible!**
The order of mixing couldn't possibly affect the taste... could it?

The Challenge

“Prove it.”

The room fell silent. How could they test such a claim?

- If she guesses correctly once, is that proof? *(Could be luck...)*
- What if she gets 2 right? 3 right? *(Still could be luck...)*
- How many correct answers would **convince** them she has real ability?

This is the fundamental question of statistics:

How do we distinguish real effects from random chance?

Enter: Ronald Fisher

One man at the table saw this as more than a parlor game.

Ronald Fisher – a young statistician working at the Rothamsted Experimental Station – realized this simple question about tea contained a profound scientific problem.

“I can design an experiment to test this,” he said.

What followed would revolutionize science forever.

Who Was Ronald Fisher?

The "Father of Modern Statistics"

- Born in London, 1890
- Studied mathematics at Cambridge
- Poor eyesight prevented WWI service
- Worked on agricultural experiments
- Revolutionized scientific methodology

His Contributions:

- **P-value** – probability under null
- **Null hypothesis** – default assumption
- **ANOVA** – comparing groups
- **Maximum likelihood estimation**
- **Experimental design principles**

"To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination." – R.A. Fisher

Fisher's Elegant Experiment

The Design:

1. Prepare **8 cups** of tea: 4 with milk first, 4 with tea first
2. Present them in **random order**
3. Tell the lady there are exactly 4 of each type
4. Ask her to identify which 4 cups had milk added first

The Mathematics:

- Total ways to choose 4 from 8: $\binom{8}{4} = \frac{8!}{4! \times 4!} = 70$
- Only **1 way** to get all 4 correct
- $P(\text{perfect by chance}) = \frac{1}{70} = 1.4\%$

Fisher's Reasoning:

If she gets all 4 correct, there's only a 1.4% chance she was guessing.

This is small enough to **reject** the idea that she's just lucky!

The Birth of Hypothesis Testing

Fisher formalized this into a framework:

1. **Null Hypothesis (H_0):** The lady has no ability (just guessing)
2. **Alternative Hypothesis (H_1):** The lady has real ability
3. **P-value:** Probability of the result if H_0 is true
4. **Decision:** If p-value < 0.05, reject H_0

The Result: Dr. Bristol correctly identified **all 8 cups!**

$$\text{P-value} = 1.4\% < 5\%$$

She really could taste the difference.

From Tea Cups to Heart Failure

The exact same logic applies to our analysis:

Lady Tasting Tea	**Heart Failure Analysis**
Can she distinguish milk-first from tea-first?	Can serum creatinine distinguish survivors from non-survivors?
H_0: She's guessing randomly	H_0: No difference between groups
If $p < 0.05 \rightarrow$ real ability	If $p < 0.05 \rightarrow$ feature matters

Fisher's Legacy:

- These methods are now used in **medicine**, **biology**, **psychology**, and **data science**
- Every "$p < 0.05$" you see in research traces back to Fisher

What is Correlation?

Pearson Correlation measures the **linear relationship** between two variables.

Range: -1 **to** $+1$

- $+1$: Perfect positive (both increase together)
- 0: No linear relationship
- -1: Perfect negative (one up, other down)

Formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Key Question: Which features are correlated with death?

Feature Correlations with DEATH_EVENT

Positive Correlations:

- serum_creatinine: +0.29
- age: +0.25

Higher values → higher death risk

Negative Correlations:

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

Higher values → lower death risk

Key Insight:
Follow-up time has the strongest correlation – but this is expected!

Patients who die have shorter follow-up periods.

Warning: 'time' is a "leaky" feature – we wouldn't know it in advance for prediction!

T-Test: Comparing Group Means

Question: Is there a significant difference in feature values between survivors and non-survivors?

How it works:

1. Calculate mean for each group
2. Measure the difference
3. Account for variability
4. Get a p-value

Assumptions:

- Normal distribution
- Independent samples
- (Welch's t-test relaxes equal variance)

Null Hypothesis: No difference between group means ($\mu_1 = \mu_2$)

Mann-Whitney U Test: Non-Parametric Alternative

When to use: Data is NOT normally distributed

How it works:

1. Rank all values together
2. Sum ranks for each group
3. Compare rank sums

Advantages:

- No normality assumption
- Robust to outliers
- Works with ordinal data

Example:

Age values: [55, 60, 65, 70, 75]
Ranks: [1, 2, 3, 4, 5]

If survivors have mostly low ranks and non-survivors have high ranks, there's a significant difference.

Significant Features ($p < 0.05$)

Feature	T-test p-value	Mann-Whitney p-value	Significant?
time	2.3×10^{-22}	6.9×10^{-21}	Yes
ejection_fraction	9.6×10^{-6}	7.4×10^{-7}	Yes
serum_creatinine	6.4×10^{-5}	1.6×10^{-10}	Yes
age	4.7×10^{-5}	1.7×10^{-4}	Yes
serum_sodium	1.9×10^{-3}	2.9×10^{-4}	Yes
diabetes	0.97	0.97	No
sex	0.94	0.94	No
smoking	0.83	0.83	No

The Problem with Multiple Testing

Scenario: Testing 12 features at $\alpha = 0.05$

The Math:

- Each test: 5% chance of false positive
- 12 tests: Expected false positives $= 12 \times 0.05 = 0.6$
- Probability of at least one false positive: $1 - (0.95)^{12} = 46\%$

Real-World Problem:

- Publish 20 studies
- 1 will show "significant" result by chance
- This is why many studies don't replicate!

Solution: Adjust p-values to control the False Discovery Rate (FDR)

Benjamini-Hochberg (FDR) Correction

Goal: Control the expected proportion of false discoveries

Procedure:

1. Rank p-values from smallest to largest: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$
2. For each p-value, calculate the adjusted value:

$$p_{adj} = p \times \frac{m}{\text{rank}}$$

3. Compare adjusted p-values to α

Why FDR over Bonferroni?

- Bonferroni: Very conservative (α/m) – may miss real effects
- FDR: Allows some false positives, but controls the rate

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

Conclusion: Our 5 significant features remain significant even after accounting for multiple testing!

Statistical Tests vs. Feature Importance

Statistical Tests	Feature Importance
"Is there a difference between groups?"	"How useful for prediction?"
Tests one feature at a time	Considers all features together
Measures statistical significance	Measures predictive power
Needs assumptions (normality, etc.)	Often model-based

Key Insight: A feature can be statistically significant but not useful for prediction (and vice versa). We need BOTH perspectives!

What is Random Forest?

Random Forest = A collection of many decision trees that "vote" together

Analogy:

Like asking 100 doctors for their opinion:

- Each doctor (tree) looks at different aspects
- They all vote on the outcome
- Majority vote wins

Why Random Forest?

- Automatically learns useful features
- Captures non-linear relationships
- Robust and widely used
- Provides feature importance for free!

Gini Importance (Mean Decrease in Impurity)

How it works:

1. At each split, the tree asks: “Which feature best separates survived vs. died?”
2. Features that create better splits are used more often
3. Importance = how much each feature reduces “impurity” (mixing of classes)

Simple Example:

- If `ejection_fraction < 30` perfectly separates patients → HIGH importance
- If `smoking` doesn’t help separate groups → LOW importance

Limitation: Can be biased toward features with many unique values

Permutation Importance: A Better Alternative

Problem with Gini: Biased toward features with many values

Permutation Importance Solution:

1. Train model and measure accuracy
2. Randomly shuffle one feature's values
3. Measure how much accuracy drops
4. Bigger drop = more important feature

Intuition: If shuffling a feature hurts the model a lot, that feature was important!

Advantages:

- Unbiased
- Works with any model
- Evaluated on test data (more reliable)

Feature Importance Results

Top 5 Most Important Features:

1. **time** – Follow-up period (but “leaky”!)
2. **serum_creatinine** – Kidney function indicator
3. **ejection_fraction** – Heart pumping efficiency
4. **age** – Patient age
5. **serum_sodium** – Electrolyte balance

Clinical Interpretation:

- **Serum creatinine**: High levels indicate kidney dysfunction
- **Ejection fraction**: Low values mean heart isn’t pumping efficiently
- These match the original research paper findings!

Why Does Feature Variance Matter?

Core Idea: A feature that doesn't vary can't help distinguish groups!

Example:

- If ALL patients have `diabetes = 1`
- This feature tells us nothing
- Can't distinguish survivors from non-survivors

Variance Formula:

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Measures spread around the mean

Rule: Remove features with very low variance (they carry no information)

Coefficient of Variation (CV)

Problem: Raw variance depends on scale

- Platelets: variance $= 9.5 \times 10^9$ (large numbers!)
- Anaemia: variance $= 0.25$ (binary $0/1$)
- Can't compare directly!

Solution: Coefficient of Variation

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} = \frac{\sigma}{\mu}$$

- Scale-free measure of relative spread
- CV > 1: High variability relative to mean
- CV < 1: Low variability relative to mean

What is Multicollinearity?

Definition: When two or more features are highly correlated

Simple Example:

- Height in cm
- Height in inches
- Same information!

In Our Data:

- sex and smoking: $r = 0.45$
- (Men smoke more in this dataset)

Why is it a problem?

- Redundant information
- Hard to interpret importance
- Unstable regression coefficients
- Wastes model capacity

Variance Inflation Factor (VIF)

VIF measures how much a feature's variance is "inflated" by correlation with others

How it works:

1. For feature X_j, predict it using ALL other features
2. Calculate R_j^2 (how well others predict X_j)
3. VIF formula: $\text{VIF}_j = \frac{1}{1-R_j^2}$

Intuition:

- If X_j predicted perfectly by others ($R^2 = 1$): $\text{VIF} \to \infty$
- If X_j is independent ($R^2 = 0$): $\text{VIF} = 1$

VIF Interpretation Guide

VIF Value	**Interpretation & Action**
$\text{VIF} = 1$	No correlation – no action needed
$1 < \text{VIF} < 5$	Moderate – usually acceptable
$5 \leq \text{VIF} < 10$	High – investigate further
$\text{VIF} \geq 10$	Severe – consider removing feature

What to do with high VIF?

- Remove one of the correlated features
- Combine features (e.g., PCA)
- Use regularization (Ridge, Lasso)

VIF in Python

```
from statsmodels.stats.outliers_influence
     import variance_inflation_factor
from sklearn.preprocessing import StandardScaler

# Scale features for numerical stability
X_scaled = StandardScaler().fit_transform(X)

# Calculate VIF for each feature
vif_data = pd.DataFrame()
vif_data[’Feature’] = X.columns
vif_data[’VIF’] = [
    variance_inflation_factor(X_scaled, i)
    for i in range(X_scaled.shape[1])
]
vif_data.sort_values(’VIF’, ascending=False)
```

VIF Results for Our Dataset

Feature	VIF	Status
sex	1.35	OK
smoking	1.32	OK
age	1.15	OK
time	1.14	OK
serum_sodium	1.13	OK
ejection_fraction	1.10	OK
serum_creatinine	1.08	OK

Good news!

- All VIF values are close to 1
- No severe multicollinearity
- All features provide relatively independent information

Note: sex and smoking have slightly higher VIF due to their correlation (0.45)

Variance vs. VIF: Key Differences

Variance	**VIF**
Measures spread of a *single* feature	Measures correlation *between* features
Low variance = feature doesn't vary much	High VIF = feature is redundant
Question: "Does this feature have different values?"	Question: "Is this feature's info already captured by others?"
Solution: Remove low-variance features	Solution: Remove one of correlated pair

Key Takeaways

1. **Correlation** measures linear relationships between variables
2. **T-test** compares means (assumes normality)
3. **Mann-Whitney U** is non-parametric (no normality assumption)
4. **FDR Correction** controls false discovery rate in multiple testing
5. **Random Forest** provides feature importance scores
6. **Permutation Importance** is more reliable than Gini importance
7. **Variance/CV** tells us how much features vary
8. **VIF** detects multicollinearity (redundant features)

Key Findings for Heart Failure Dataset

Most Important Predictive Features:

1. time (but leaky!)
2. serum_creatinine
3. ejection_fraction
4. age
5. serum_sodium

Data Quality:

- No multicollinearity issues (all VIF < 5)
- All features have adequate variance
- Results match the original research paper!

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. Compare feature rankings from t-test vs. Random Forest
4. Calculate VIF for all features and interpret results
5. Write a low-variance filter function

Resources:

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