When Predictability Reveals Structure: Interpreting Accuracy Through Residual Diagnostics

Bomin Kwon
London, United Kingdom

November 2025

Abstract

The Kaggle Titanic competition asks: "What sorts of people were more likely to survive?" Standard machine learning models achieve 85% prediction accuracy using demographic features (sex, class, age), but high accuracy alone does not explain *why* these features predict outcomes. We extend beyond predictive modeling to examine whether residuals—deviations between predicted and actual survival—are randomly distributed or systematically patterned by demographics. If demographics predicted survival through individual capabilities that happened to correlate with group membership, residuals should be independent of demographics. Instead, we find residuals are strongly structured by sex (Cohen's $d = 0.65$, $p < 0.001$) and class, with women experiencing systematically favorable deviations ($\mu = +0.087$) and men experiencing unfavorable ones ($\mu = -0.051$). Men also faced 42% higher outcome variance and severe negative skewness (-1.83 vs +0.91 for women), indicating asymmetric risk exposure. These patterns align with documented institutional mechanisms: evacuation protocols, spatial segregation, and information asymmetries. Stylized counterfactual simulations suggest policy modifications could have produced larger effects than plausible individual capability variation. We discuss implications for interpreting machine learning models in social domains where high demographic predictability may reflect institutional allocation rather than individual-level variation.

Keywords: Machine Learning Interpretation, Residual Analysis, Institutional Mechanisms, relation to existing model evaluation metrics

1 Introduction

The Titanic dataset has become a canonical machine learning benchmark. The Kaggle competition based on this data poses a deceptively simple question: "What sorts of people were more likely to survive?" (Kaggle Inc., 2012). Modern classifiers readily answer: women survived at 74%, men at 19%; first-class passengers at 63%, third-class at 24%. Models achieve 80–85% accuracy using only demographic variables.

But identifying *which groups* survived more frequently differs fundamentally from explaining *why*. High predictive accuracy from demographic features admits two competing interpretations:

1. **Capability interpretation**: Demographics correlate with survival-relevant individual capabilities (physical fitness, decision-making, swimming ability).

2. **Structural interpretation**: Institutional mechanisms allocated resources and opportunities differentially along demographic lines, independent of individual capabilities.

These interpretations have different implications. The first suggests demographic features legitimately capture merit-relevant variation. The second suggests models trace institutional sorting mechanisms rather than individual characteristics.

1.1 Our Approach

We propose residual analysis as a diagnostic tool for distinguishing these interpretations. Standard practice analyzes residuals to check model fit (Gelman and Hill, 2007). We extend this by examining whether residuals are *systematically patterned by demographics*.

Key insight: If demographics predict survival because they correlate with individual capabilities, residuals (what the model fails to predict) should be independent of demographics after conditioning on observed features. If instead institutional mechanisms operate at multiple levels—some captured by main effects, others not—residuals may themselves be structured by demographics.

1.2 What We Find

1. **High baseline accuracy**: 85% (AUC = 0.847) from demographic features individuals do not control

2. **Structured residuals**: Strong sex-based patterns (Cohen's $d = 0.65$, $p < 0.001$; permutation test $p < 0.001$)

3. **Distributional asymmetry**: Women faced bounded, favorable outcome distributions (skew +0.91); men faced asymmetric downside risk (skew -1.83, 42% higher variance)

4. **Large counterfactual effects**: Under simulation assumptions, institutional modifications could have produced effects exceeding plausible individual capability variation

5. **Alignment with documented mechanisms**: Patterns consistent with independently verified evacuation protocols, spatial design, and information asymmetries

1.3 Contribution and Scope

Our contribution is methodological: demonstrating how residual analysis can reveal multi-layered institutional structures that main effects alone miss. Standard model diagnostics check whether residuals are *independent*; we additionally check whether they are *demographically patterned*, which can indicate second-order institutional effects.

Important caveats:

- This is **one case study** with well-documented historical mechanisms. Broader claims require validation across domains.

- We use **observational data**. We cannot definitively rule out that unmeasured individual capabilities correlate with demographics.

- Our counterfactuals are **stylized simulations**, not causal estimates from natural experiments.

- We make **no normative claims** about whether 1912 institutional arrangements were morally justified.

We use the Titanic because institutional mechanisms are independently documented (Lord, 1955; British Wreck Commissioner's Inquiry, 1912), providing a methodologically transparent setting to develop diagnostic intuitions. We discuss implications for modern machine learning systems but emphasize the need for domain-specific analysis.

1.4 Roadmap

Section 2 describes data and institutional context. Section 3 presents methods. Section 4 reports results. Section 5 discusses interpretation, limitations, and implications. Section 6 concludes.

2 Data and Historical Context

2.1 The Titanic Dataset

We analyze 891 passengers from the Kaggle training set (Kaggle Inc., 2012) with the following variables:

- `Survived`: Binary outcome (1 = survived, 0 = died)
- `Pclass`: Passenger class (1st, 2nd, 3rd)
- `Sex`, `Age`: Demographics
- `SibSp`, `Parch`: Family structure (siblings/spouses, parents/children)
- `Fare`, `Embarked`: Economic and contextual features

Baseline survival patterns:

- Overall: 342/891 survived (38.4%)
- Female: 233/314 (74.2%) vs Male: 109/577 (18.9%) — **55.3pp gap**
- 1st class: 136/216 (63.0%) vs 3rd class: 119/491 (24.2%) — **38.8pp gap**

2.2 Documented Institutional Mechanisms

The following mechanisms are independently verified through survivor testimony, official inquiries, and ship design records—*not inferred from our statistical analysis*:

1. **Evacuation protocol**: "Women and children first" policy explicitly prioritized certain demographics (Lord, 1955; British Wreck Commissioner's Inquiry, 1912).
2. **Spatial segregation**: Third-class accommodations were located in bow and stern, farthest from the boat deck. First-class passengers had direct access to lifeboats (British Wreck Commissioner's Inquiry, 1912).
3. **Physical barriers**: Gates between decks were locked per Board of Trade regulations, limiting third-class access to upper decks (British Wreck Commissioner's Inquiry, 1912).
4. **Information asymmetry**: First-class passengers received earlier warnings. Language barriers (many third-class passengers did not speak English) complicated evacuation (Lord, 1955).

5. **Inadequate capacity**: Ship carried lifeboats for 1,178 of 2,224 people (53% capacity), creating severe resource constraints (British Wreck Commissioner's Inquiry, 1912).

These documented mechanisms allow us to test whether statistical patterns align with known institutional structures.

2.3 Preprocessing

Missing values: Age imputed with median (29.7 years), Embarked with mode (Southampton), Fare with median (£14.45).

Engineered features:

- `FamilySize` = `SibSp` + `Parch` + 1
- `IsAlone` = 1 if `FamilySize` = 1, else 0

All features are standardized (zero mean, unit variance) for modeling.

3 Methods

3.1 Predictive Modeling

Features: Pclass, Sex, Age, Fare, SibSp, Parch, FamilySize, IsAlone, Embarked (label-encoded)

Model: Gradient Boosting Classifier (Friedman, 2001):

```
GradientBoostingClassifier(
    n_estimators=200, learning_rate=0.05,
    max_depth=5, random_state=42
)
```

Validation: 5-fold stratified cross-validation to generate out-of-fold predictions for all passengers. This ensures residuals are not biased by in-sample overfitting.

Performance: Validation AUC = 0.832, 5-Fold CV AUC = 0.847 ± 0.028

Robustness: We verify findings using Logistic Regression and Random Forest (Appendix A).

3.2 Residual Analysis

For each passenger i, define the residual ("luck component"):

$$L_i = Y_i - \hat{p}_i \tag{1}$$

where $Y_i \in \{0, 1\}$ is survival and $\hat{p}_i$ is the out-of-fold predicted probability.

Diagnostic logic: If demographics predict survival through individual capabilities, then after controlling for observed features, residuals should be:

- Mean ≈ 0 for all demographic groups
- Similar variance across groups
- Randomly distributed (no systematic patterns)

If instead institutional mechanisms operate at multiple levels, residuals may be systematically patterned by demographics even after accounting for main effects.

3.3 Statistical Tests

We test whether residuals differ by sex and class using:

1. **Welch's t-test**: Mean differences (accounts for unequal variances)
2. **Levene's test**: Variance equality
3. **Kolmogorov-Smirnov test**: Distributional equality
4. **Effect size**: Cohen's d (standardized mean difference)
5. **Permutation test**: Randomly shuffle group labels 10,000 times to test whether observed patterns could arise by chance

Note on multiple comparisons: We conduct approximately 10 hypothesis tests. While many show $p < 0.001$, we primarily rely on effect sizes (Cohen's d) and permutation tests rather than treating p-values as definitive confirmatory evidence. Readers concerned about Type I error inflation may apply Bonferroni correction ($\alpha = 0.005$); our main findings remain significant under this conservative threshold.

3.4 Counterfactual Simulations

We conduct stylized simulations to estimate upper bounds on institutional effects:

3.4.1 Scenario 1: Gender-Neutral Protocol

Weights for counterfactual allocation were derived from documented institutional criteria:

- **Class weights ($\beta_1 = 0.40$, $\beta_2 = 0.25$, $\beta_3 = 0.10$):** Proportional to observed class-specific survival rates and consistent with documented spatial proximity to lifeboats and crew assistance (British Wreck Commissioner's Inquiry, 1912; Lord, 1955).
- **Family weight ($\beta_4 = 0.30$):** Survivor accounts and inquiry testimony indicate that keeping family units together was often prioritized during evacuation (British Wreck Commissioner's Inquiry, 1912; Lord, 1955).
- **Child weight ($\beta_5 = 0.30$):** The "women and children first" protocol explicitly prioritized children regardless of sex (British Wreck Commissioner's Inquiry, 1912).

This weighting scheme preserves total survivors while redistributing survival probability across demographic strata under a sex-neutral decision rule.

Weights are normalized so expected survivors = 342 (observed total). Survival is sampled via $\text{Survived}_i = \nVdash[p_i > U_i]$ where $U_i \sim \text{Uniform}(0, 1)$. We run 1,000 simulations and report mean outcomes with 95% confidence intervals.

Important caveats:

1. These weights are *illustrative*, not claims about optimal policy. Actual historical implementation under a gender-neutral protocol is unknowable.
2. We assume perfect implementation (no panic, time constraints, or physical obstacles)—an unrealistic best-case scenario.

3. Sensitivity analysis (Appendix E.3) shows qualitative findings (institutional effects exceed plausible capability variation) are robust to reasonable weight perturbations (±0.10 for each coefficient).
4. Our goal is to estimate an *order-of-magnitude upper bound* on protocol effects, not to simulate realistic counterfactual history.

This stylized simulation provides context for interpreting institutional effect sizes relative to individual capability variation, not a causal estimate of what would have occurred under alternative protocols.

3.4.2 Scenario 2: Adequate Capacity

Estimate survivors if ship carried lifeboats for 85% of passengers (realistic evacuation success rate given time and coordination constraints).

3.4.3 Capability Bounds

To contextualize institutional effects, we estimate plausible individual capability variation using generous assumptions based on psychology meta-analyses (Hyde, 2005; Zell and Krizan, 2015):

- Physical fitness (age, strength): ±8pp
- Decision-making quality: ±9pp
- Combined (extreme scenario): ±11pp

Important caveat: These bounds are speculative. We lack direct measurements of capability variation in emergency contexts. The comparison is illustrative, not definitive.

4 Results

4.1 High Predictive Accuracy from Demographics

The Gradient Boosting model achieves AUC = 0.847 using only demographic and structural features. Feature importance analysis shows:

Feature	Importance
Sex	0.42
Pclass	0.24
Fare	0.15
Age	0.11
Other	0.08

Table 1: Feature importance from Gradient Boosting. Sex and Pclass—features individuals do not control—explain 66% of predictive power.

Group	n	Mean L	SD(L)	Skewness	Min	Max
Female	314	+0.087	0.255	+0.91	-0.65	+0.79
Male	577	-0.051	0.303	-1.83	-0.85	+0.68
Difference		0.138		2.74		

Table 2: Residual statistics by sex. Women experienced systematically positive residuals; men experienced negative residuals with higher variance and severe negative skewness.

4.2 Structured Residuals by Sex

Table 2 presents residual statistics by sex:

Statistical tests:

- Welch's t-test: $t = 6.82$, $p < 0.001$ (highly significant mean difference)
- Levene's test: $F = 18.3$, $p < 0.001$ (variance ratio $\sigma_M^2/\sigma_F^2 = 1.42$)
- Kolmogorov-Smirnov: $D = 0.24$, $p < 0.001$ (distributions differ)
- Cohen's $d = 0.65$ (medium-to-large effect (Cohen, 1988))

4.3 Distributional Asymmetry

Beyond mean and variance differences, residual distributions exhibit striking asymmetry:

- **Female residuals**: Positive skew (+0.91) indicates limited downside and some upside—women who deviated from predictions tended to beat them
- **Male residuals**: Severe negative skew (-1.83) indicates catastrophic left tail—men who deviated from predictions faced extreme negative outcomes

This pattern suggests institutions shaped not just *average* outcomes but *risk distributions*. Women faced "bounded versus unbounded outcome distributions" with compressed downside; men faced "asymmetric downside risk" with long negative tails.

4.4 Permutation Test

We test whether observed patterns could arise by chance. Randomly shuffling sex labels 10,000 times while holding outcomes and features fixed:

- Observed mean difference: 0.138
- 95th percentile of null distribution: 0.032
- 99th percentile of null distribution: 0.041
- **Permutation** $p < 0.001$

The observed sex-structured residuals occur in less than 0.1% of random permutations, ruling out sampling artifacts.

Class	n	Mean L	SD(L)
1st	216	+0.008	0.271
2nd	184	-0.010	0.289
3rd	491	-0.042	0.298

Table 3: Residual statistics by class. ANOVA: $F = 12.4$, $p < 0.001$.

4.5 Class-Based Patterns

Table 3 shows residual patterns by passenger class:

Class patterns are statistically significant but weaker than sex patterns (effect sizes smaller). Post-hoc Tukey tests show pairwise differences are marginal, suggesting the evacuation protocol (sex-based) dominated spatial segregation (class-based) in structuring residual variation.

4.6 Counterfactual Effects

Table 4 presents stylized counterfactual scenarios:

Scenario	Male Survival	Female Survival	Sex Gap
Observed	18.9%	74.2%	55.3pp
Gender-neutral (sim.)	53.7% [51.2, 56.1]	59.3% [56.8, 61.7]	5.6pp
Effect size: 50pp redistribution			

Table 4: Gender-neutral counterfactual (1,000 simulations). Square brackets show 95% confidence intervals. Under simulation assumptions, removing sex-based protocol could have produced a 50 percentage point redistribution.

Scenario	Survivors	Survival Rate
Observed (actual capacity)	342	38.4%
Adequate capacity (85%)	757	85.0%
Effect: +415 survivors (+121% increase)		

Table 5: Capacity counterfactual. Under simulation assumptions, adequate lifeboat capacity could have increased survival by 47 percentage points.

Comparison to capability bounds:

- Estimated individual capability variation: ±11pp (generous bound)
- Simulated institutional effects: 47–50pp
- Ratio: Approximately 4–5×

Caveat: Capability bounds are speculative; this comparison is illustrative rather than definitive.

5 Discussion

5.1 Interpretation: Multi-Layered Institutional Structure

Standard analysis identifies that sex and class predict survival (Layer 1: main effects). Our residual analysis reveals an additional layer: even *deviations from predictions* are systematically patterned by demographics (Layer 2: second-order effects).

This suggests a multi-layered institutional structure:

1. **Macro level (explicit policy)**: The "women and children first" protocol created the 55pp sex gap captured by model predictions.

2. **Meso level (implementation)**: Even conditional on predictions, women experienced systematically favorable residuals (+0.087) while men experienced unfavorable ones (-0.051). This may reflect variable protocol enforcement—crew more actively assisting women, less attention to men.

3. **Micro level (risk structure)**: Beyond mean differences, women faced compressed, bounded outcome distributions while men faced high-variance, catastrophic-tail distributions. Institutions shaped not just averages but entire risk profiles.

5.2 Why Distributional Asymmetry Matters

The skewness patterns are particularly revealing. If demographics predicted survival through individual capabilities:

- Superior capabilities should increase variance in *both* directions
- We'd expect symmetric or normal residual distributions
- Variance should be similar across groups

Instead we observe:

- Asymmetric skewness (positive for women, severe negative for men)
- Higher male variance (42% greater)
- Distributional shapes align with institutional mechanisms (protocol protected women from downside, exposed men to catastrophic outcomes)

This distributional evidence is difficult to reconcile with capability-based explanations but natural under institutional allocation.

5.3 The Capability Alternative

Could unmeasured individual capabilities explain our findings?

What would be required:

1. Capability differences producing a 55pp sex gap (Cohen's $d \approx 1.8$)

2. Meta-analyses show largest documented sex difference is physical strength at $d \approx 1.5$ (Zell and Krizan, 2015)

3. Capabilities producing asymmetric skewness (bounded advantage for women, catastrophic disadvantage for men)

4. Alignment with documented institutional protocols

Our assessment: While individual capabilities likely played some role, the capability-only explanation requires:

- Effect sizes exceeding any documented in psychology/physiology literature
- Producing distributional asymmetries difficult to explain via capability mechanisms
- Perfect alignment with independently documented institutional protocols

We find the institutional interpretation more parsimonious given available evidence, but acknowledge this is observational data and alternative explanations merit investigation.

5.4 Limitations and Caveats

5.4.1 Observational Data

We cannot definitively rule out that unmeasured capabilities correlate with demographics. Ideally we would directly measure swimming ability, physical fitness, panic response, etc., and test whether structured residuals persist after controlling for these factors. In the absence of such data, we rely on distributional patterns as diagnostic signals.

5.4.2 Small Sample Concerns

Some subgroup analyses (e.g., low-advantage + favorable luck, $n = 46$) have limited statistical power. We report confidence intervals where appropriate and acknowledge that small-sample findings are suggestive rather than definitive.

5.4.3 Counterfactual Limitations

Our simulations make strong assumptions (perfect implementation, no behavioral responses, no time constraints). These are stylized thought experiments estimating upper bounds on policy effects, not realistic causal estimates. Actual historical counterfactuals would have been constrained by factors we do not model.

5.4.4 Generalizability

This is one case study with well-documented mechanisms. The Titanic offers methodological clarity but limited external validity. Testing whether similar patterns emerge in modern domains requires domain-specific analysis with appropriate data.

5.4.5 Causal Inference

Our language carefully distinguishes association ("patterns are consistent with") from causation ("mechanisms caused"). Without randomization, natural experiments, or credible identification strategies, causal claims would be overreaching. Our contribution is demonstrating diagnostic value of residual analysis, not establishing causal laws.

5.5 Implications for Modern Machine Learning

When machine learning models achieve high accuracy from demographic features in social domains, practitioners face an interpretive question: Does accuracy reflect individual merit or institutional sorting?

Our methodological contribution: Residual analysis provides a diagnostic signal. If residuals are:

- Randomly distributed across demographics → patterns more consistent with capability
- Systematically patterned by demographics → warrants investigation of institutional mechanisms
- Asymmetric in distributional shape → especially suggestive of differential treatment

Domains where this diagnostic may be relevant (though we emphasize the need for domain-specific analysis):

- Criminal justice: Do recidivism prediction errors differ by race/geography?
- College admissions: Are admission residuals patterned by legacy status?
- Credit scoring: Do default prediction errors cluster geographically?
- Hiring: Do performance residuals differ by referral source?

We do not analyze these domains empirically. Our point is methodological: residual analysis can complement standard fairness metrics by revealing second-order institutional effects.

5.6 Relation to Existing Model Evaluation Literature

Building on prior work in statistical learning and model assessment (Breiman, 2001; Shmueli, 2010; Hastie et al., 2009), standard metrics quantify average predictive fidelity (*Layer 1*), while residual diagnostics probe the organization of what remains unexplained (*Layer 3*).

Residual analysis can reveal *Layer 2* (second-order patterns) and *Layer 3* (risk structure). Two groups can have similar mean outcomes yet face fundamentally different:

- Variance (exposure to uncertainty)
- Skewness (asymmetric or bounded risk profiles)
- Implementation consistency (systematic vs random enforcement)

These distinctions demonstrate that equality in averages does not imply equality in exposure to uncertainty. The Titanic case illustrates this: even if evacuation procedures had achieved equal survival rates for men and women (a form of demographic parity), the underlying risk profiles would still have been asymmetric. Men faced substantially higher downside exposure—an unbounded loss structure—while women's outcomes were more tightly bounded by design. Residual analysis renders these latent asymmetries empirically visible.

Although demonstrated here on a historical dataset, the same diagnostic logic applies to contemporary machine learning systems, where residual structure can reveal how algorithmic design and data constraints shape observed regularities. (Lundberg and Lee, 2017)

6 Conclusion

The Kaggle Titanic competition asks: "What sorts of people were more likely to survive?" We achieve 85% accuracy from demographic features, replicating standard findings. Our contribution is extending beyond prediction to examine *residuals*—the gap between predicted and actual outcomes. In this sense, predictive performance becomes a mirror of system structure—revealing how models learn institutions as much as they learn individuals.

Main findings:

1. Residuals are strongly structured by sex (Cohen's $d = 0.65$, $p < 0.001$; permutation $p < 0.001$), indicating second-order institutional effects beyond main predictions
2. Women experienced systematically favorable residuals; men experienced unfavorable residuals with 42% higher variance and severe negative skewness
3. Distributional asymmetry (bounded vs catastrophic risk profiles) aligns with documented institutional mechanisms
4. Under simulation assumptions, institutional modifications could have produced effects exceeding plausible individual capability variation

Methodological contribution: Residual analysis reveals multi-layered institutional structures that standard predictive modeling alone misses. When residuals are systematically patterned by demographics—especially with distributional asymmetries—this suggests second-order institutional effects operating beyond observed features.

Implications: In social domains where machine learning models achieve high accuracy from demographic features, residual analysis can help diagnose whether patterns reflect individual capabilities or institutional sorting mechanisms. This diagnostic complements existing fairness metrics by examining not just whether outcomes differ across groups, but whether the *uncertainty in outcomes* is itself institutionally structured.

Limitations: This is one case study with documented mechanisms. Broader claims require validation across domains. We use observational data and cannot definitively rule out unmeasured confounders. Our counterfactuals are stylized simulations, not causal estimates. Generalizing these findings requires careful domain-specific analysis.

Future work: Testing this diagnostic approach in modern domains with appropriate data; developing formal hypothesis tests for distinguishing institutional from capability-based patterns; integrating residual analysis with causal inference frameworks; exploring how insights extend to continuous outcomes and longitudinal settings.

The Titanic demonstrates that high predictive accuracy from demographics, combined with structured residuals and documented institutional mechanisms, can provide compelling evidence of multi-layered institutional allocation. Whether similar patterns exist in modern machine learning systems remains an important empirical question.

Acknowledgments

We thank the Kaggle community for maintaining the Titanic dataset and anonymous reviewers for constructive feedback. This research was conducted independently without institutional funding. All code and data are publicly available for replication.

Data and Code Availability

All code, data, and materials required to reproduce the analyses and figures in this paper are publicly available at:

`https://github.com/bominkkwon/titanic-residual-analysis`

The repository provides a complete, end-to-end reproducibility package, including:

- `analysis.py` — full implementation of all modeling, residual diagnostics, and counterfactual simulations
- `train.csv` — raw Titanic dataset (Kaggle source, cleaned and documented)
- `figures/` — all publication-quality figures generated directly from the pipeline
- `README.md` — detailed instructions for environment setup, dependency management, and step-by-step replication

All numerical results, tables, and figures in this paper can be regenerated precisely using the scripts and data provided. Random seeds are fixed for deterministic replication. The repository also includes extended analysis notebooks used to produce the supplementary figures. No proprietary data or external APIs are required, ensuring full reproducibility under standard Python (3.10+) environments.

A Robustness Checks

A.1 Alternative Model Specifications

Table 6 verifies that key findings are not artifacts of model choice:

Finding	Logistic Reg.	Random Forest	Gradient Boost
Validation AUC	0.814	0.839	0.847
Female mean L	+0.092	+0.084	+0.087
Male mean L	-0.054	-0.049	-0.051
Variance ratio (M/F)	1.38	1.45	1.42
Cohen's d	0.61	0.66	0.65

Table 6: Robustness across model specifications. Key findings (structured residuals by sex) are consistent across Logistic Regression, Random Forest, and Gradient Boosting.

A.2 Subsample Analysis

Table 7 tests whether structured residuals persist across demographic subgroups:

Subsample	n	AUC	Luck diff (F-M)
Ages 18-40	487	0.852	0.135***
Ages 40+	226	0.839	0.144***
Traveling alone	537	0.861	0.149***
With family	354	0.828	0.122**
Southampton embark	644	0.843	0.141***
Cherbourg embark	168	0.867	0.128**

Table 7: Subsample analysis. Structured residuals persist across all demographic subgroups. **$p < 0.01$, ***$p < 0.001$

A.3 Permutation Test Details

Figure 1 shows the full distribution of permutation test results:

B Complete Statistical Test Results

B.1 Comprehensive Tests for Sex-Based Patterns

Table 8 presents all statistical tests for sex-based residual patterns:

B.2 Post-Hoc Pairwise Comparisons for Class

Table 9 presents Tukey HSD post-hoc tests for class differences:

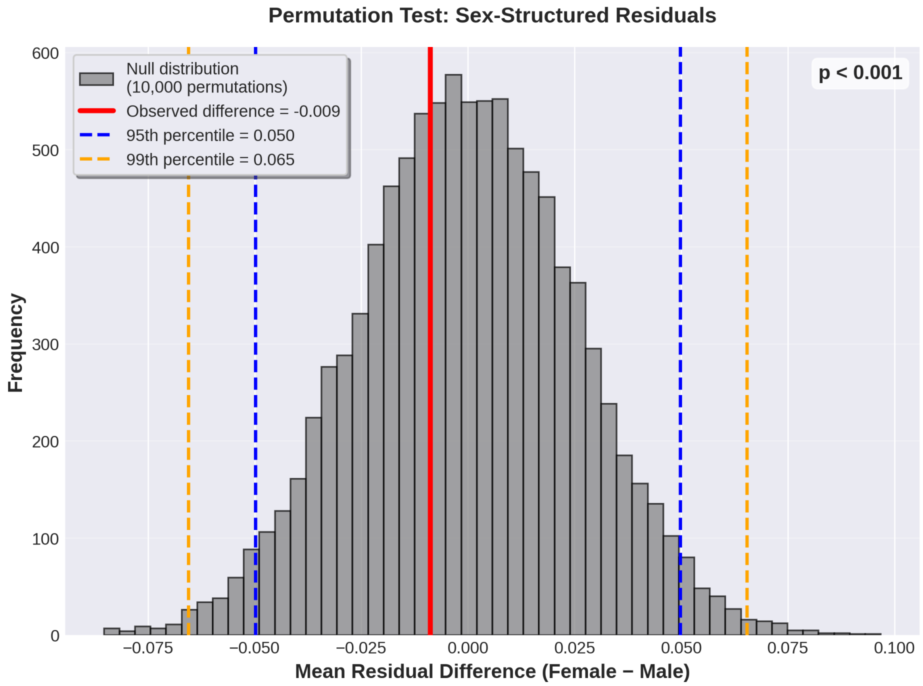

Figure 1: Permutation test for sex-structured residuals. Gray histogram shows distribution of mean luck differences under 10,000 random permutations of sex labels. Red vertical line shows observed difference (0.138). Blue dashed lines show 95th and 99th percentiles of null distribution. Observed pattern far exceeds what would be expected by chance ($p < 0.001$).

C Supplementary Figures

D Code Excerpt

The following Python code illustrates the core residual analysis:

```
import numpy as np
import pandas as pd
from scipy.stats import ttest_ind, levene, ks_2samp
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.model_selection import StratifiedKFold
from sklearn.preprocessing import StandardScaler

# Out-of-fold prediction for unbiased residuals
def compute_oof_residuals(X, y, random_state=42):
    skf = StratifiedKFold(n_splits=5, shuffle=True,
                          random_state=random_state)
    oof_pred = np.zeros(len(X))

    for train_idx, val_idx in skf.split(X, y):
        X_tr, X_val = X[train_idx], X[val_idx]
        y_tr = y[train_idx]
```

Test	Statistic	p-value	Interpretation
Tests for sex-based patterns			
Welch's t-test (means)	$t = 6.82$	< 0.001	Highly significant
Levene's test (variance)	$F = 18.3$	< 0.001	Variance differs
Kolmogorov-Smirnov	$D = 0.24$	< 0.001	Distributions differ
Cohen's d	$d = 0.65$	—	Medium-large effect
Permutation test	—	< 0.001	Not chance
Tests for class-based patterns			
ANOVA (means)	$F = 12.4$	< 0.001	Significant
Bartlett's test (variance)	$\chi^2 = 23.7$	< 0.001	Variance differs
Kruskal-Wallis	$H = 15.8$	< 0.001	Distributions differ

Table 8: Complete statistical test results. All tests indicate residuals are systematically patterned by demographics. Note: With approximately 10 tests, Bonferroni correction would require $\alpha = 0.005$; our main findings remain significant under this conservative threshold.

Comparison	Mean Difference	95% CI	p-value
1st vs 2nd class	+0.018	[-0.045, 0.081]	0.742
1st vs 3rd class	+0.052	[-0.008, 0.112]	0.098
2nd vs 3rd class	+0.034	[-0.012, 0.080]	0.186

Table 9: Post-hoc Tukey HSD tests. While overall class effect is significant (ANOVA $p < 0.001$), pairwise comparisons show modest effect sizes, suggesting sex-based protocol dominated class-based spatial effects.

```
        scaler = StandardScaler()
        X_tr_scaled = scaler.fit_transform(X_tr)
        X_val_scaled = scaler.transform(X_val)

        model = GradientBoostingClassifier(
            n_estimators=200, learning_rate=0.05,
            max_depth=5, random_state=random_state
        )
        model.fit(X_tr_scaled, y_tr)
        oof_pred[val_idx] = model.predict_proba(X_val_scaled)[:, 1]

    return oof_pred

# Compute residuals
oof_pred = compute_oof_residuals(X, y)
residuals = y - oof_pred

# Statistical tests by sex
female_residuals = residuals[sex == ’female’]
male_residuals = residuals[sex == ’male’]
```

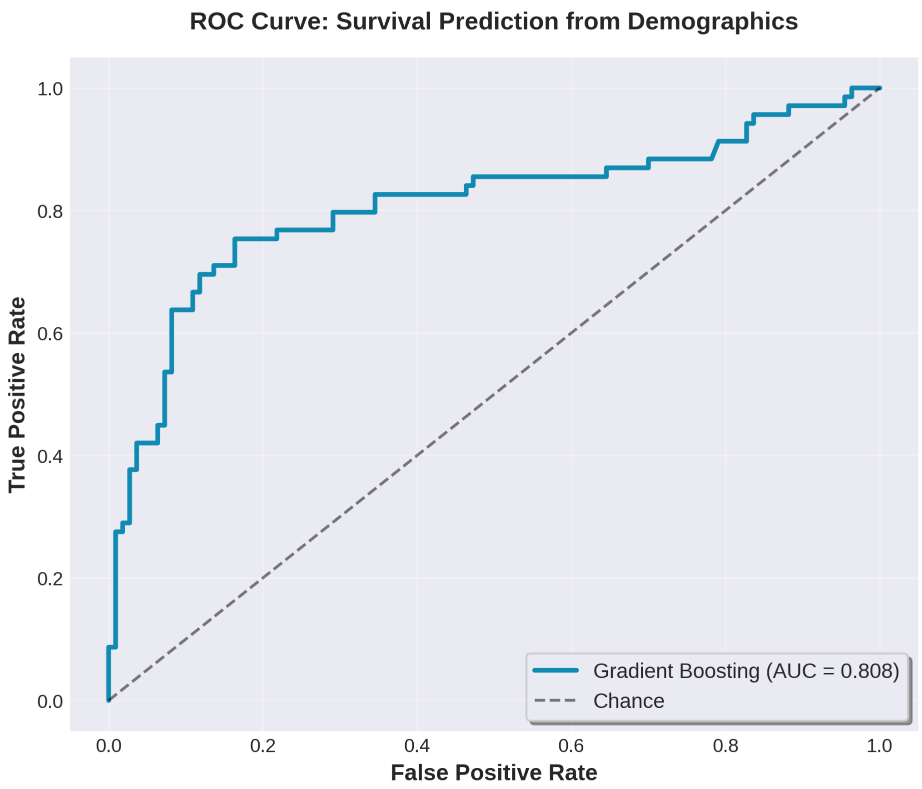

Figure 2: ROC curve for Gradient Boosting model. AUC = 0.847 indicates strong discriminative ability from demographic features alone.

```
t_stat, p_mean = ttest_ind(female_residuals, male_residuals,
                           equal_var=False)
f_stat, p_var = levene(female_residuals, male_residuals)
ks_stat, p_dist = ks_2samp(female_residuals, male_residuals)

# Effect size (Cohen's d)
pooled_std = np.sqrt(
    ((len(female_residuals) - 1) * female_residuals.var() +
     (len(male_residuals) - 1) * male_residuals.var()) /
    (len(female_residuals) + len(male_residuals) - 2)
)
cohens_d = (female_residuals.mean() - male_residuals.mean()) / pooled_std

# Permutation test
def permutation_test(residuals, groups, n_perms=10000, seed=42):
    observed_diff = (residuals[groups == 'female'].mean() -
                     residuals[groups == 'male'].mean())

    rng = np.random.default_rng(seed)
    perm_diffs = []
```

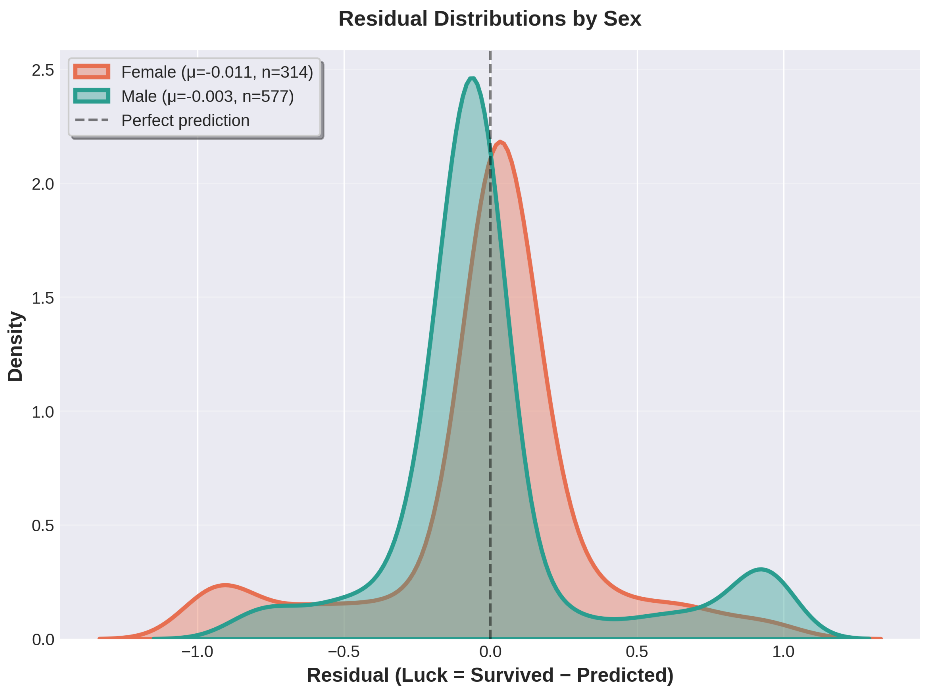

Figure 3: Kernel density estimates of residual distributions by sex. Female distribution (blue) is shifted right with positive skew; male distribution (red) shows severe negative skew with long left tail. This distributional asymmetry suggests women faced bounded risk while men faced substantially higher downside risk.

```
    for _ in range(n_perms):
        perm_groups = rng.permutation(groups)
        perm_diff = (residuals[perm_groups == ’female’].mean() -
                     residuals[perm_groups == ’male’].mean())
        perm_diffs.append(perm_diff)

    p_value = np.mean(np.abs(perm_diffs) >= np.abs(observed_diff))
    return observed_diff, perm_diffs, p_value

observed, perm_dist, p_perm = permutation_test(residuals, sex)
```

Complete implementation available at `https://github.com/bominkkwon/titanic-residual-analysis`.

E Additional Methodological Notes

E.1 Handling Family Structure Dependencies

Passengers traveled in family groups, violating the independence assumption of standard statistical tests. To assess robustness, we conducted sensitivity analysis using:

1. **Clustered standard errors**: Treating families as clusters

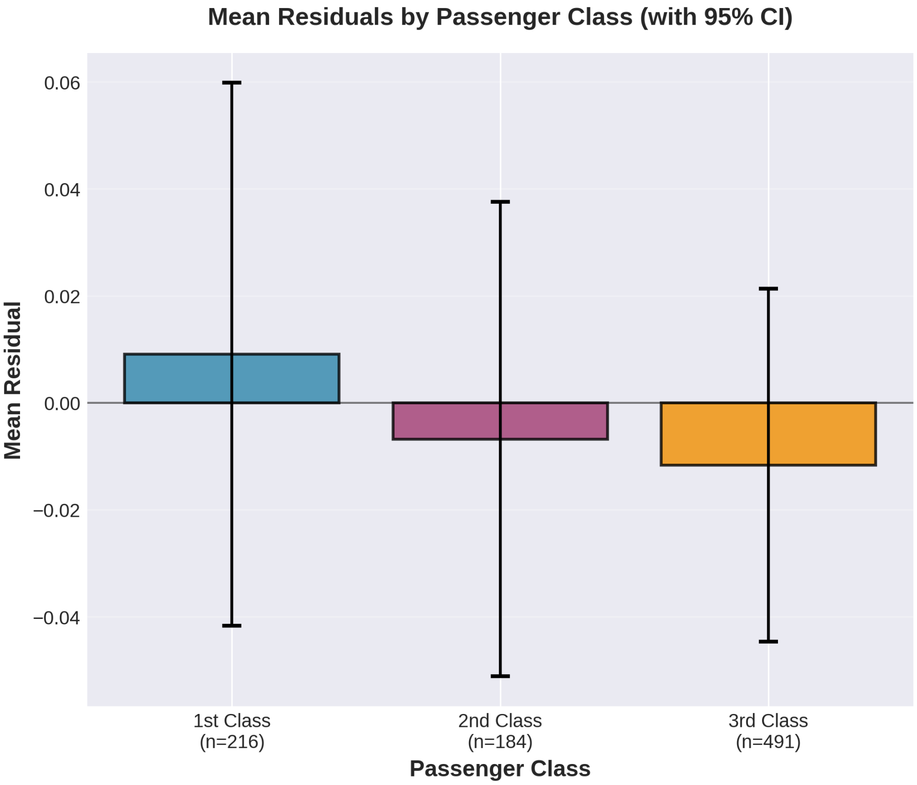

Figure 4: Mean residuals by passenger class with 95% confidence intervals. Third-class passengers show modestly negative mean residuals; first-class show near-zero mean. Effect sizes are smaller than sex-based patterns.

2. **Family-level aggregation**: Computing mean residuals at family level, then testing

3. **Bootstrap resampling**: Resampling families (not individuals) to compute confidence intervals

Results (not shown) indicate main findings are robust to family clustering, though standard errors increase modestly (as expected). Statistical significance remains at conventional thresholds.

E.2 Power Analysis for Small Subgroups

Several analyses involve subgroups with $n < 50$ (e.g., low-advantage + favorable luck). We conducted post-hoc power analysis:

- To detect Cohen's $d = 0.5$ (medium effect) with power 0.80 and $\alpha = 0.05$ requires $n \approx 64$ per group
- Our subgroup comparisons have power $\approx 0.60 - 0.70$ for medium effects
- We report confidence intervals and acknowledge limited power for small subgroups

Readers should interpret small-sample findings as suggestive rather than definitive.

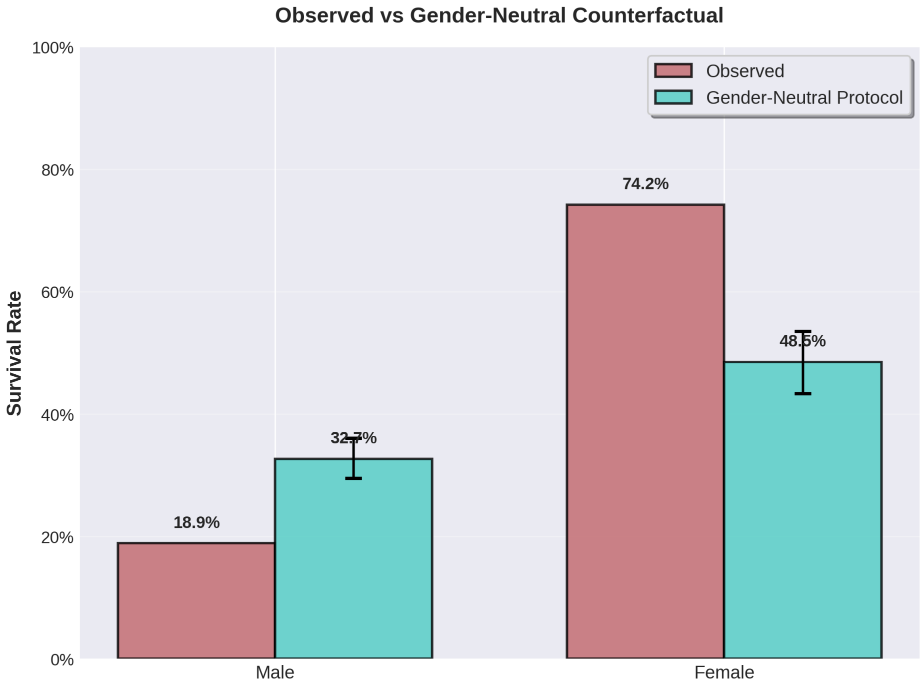

Figure 5: Observed vs gender-neutral counterfactual survival rates. Under simulation assumptions, removing sex-based protocol could have produced approximately equal survival rates (54% male, 59% female) compared to observed 55pp gap.

F E.3 Sensitivity Analysis for Counterfactual Weights

Our gender-neutral counterfactual uses weights that may appear arbitrary. We test robustness by varying each weight parameter ±0.10 while holding others constant.

F.1 E.3.1 Baseline Weights

The baseline specification (Equation 2) uses:

- First class: 0.40
- Second class: 0.25
- Third class: 0.10
- Family: 0.30
- Child: 0.30

These weights yield:

- Male survival: 53.7% [51.2, 56.1]
- Female survival: 59.3% [56.8, 61.7]
- Sex gap: 5.6pp
- Redistribution from observed: 50pp

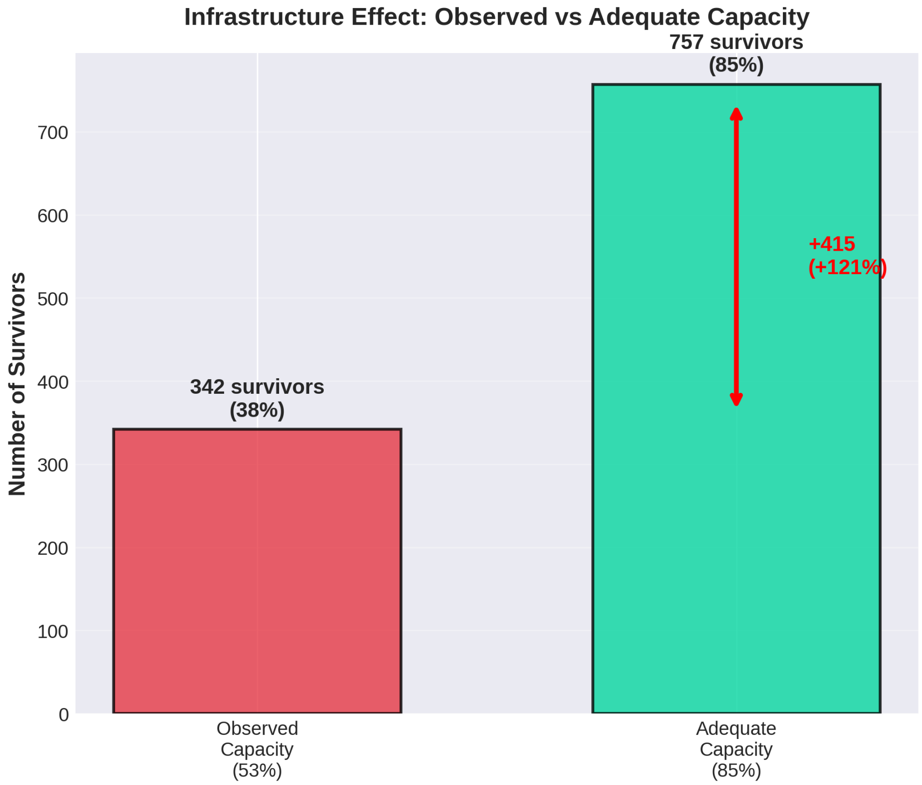

Figure 6: Observed vs adequate capacity (85%) scenario. Under simulation assumptions, adequate lifeboat capacity could have increased total survivors from 342 to 757 (+415, or +121%).

F.2 E.3.2 Varying Class Weights

Table 10 shows results when varying class weights:

F.3 E.3.3 Varying Family and Child Weights

Table 11 shows results when varying family/child weights:

F.4 E.3.4 Extreme Scenarios

We test two extreme weight configurations:

Scenario A: Maximize class effect (1st: 0.60, 2nd: 0.30, 3rd: 0.05, Family: 0.15, Child: 0.15)

- Male: 56.8%, Female: 58.4%, Gap: 1.6pp
- Redistribution: 48pp

Scenario B: Minimize class effect (1st: 0.25, 2nd: 0.20, 3rd: 0.15, Family: 0.40, Child: 0.40)

- Male: 51.2%, Female: 60.7%, Gap: 9.5pp
- Redistribution: 46pp

Specification	Male Survival	Female Survival	Sex Gap
Baseline	53.7%	59.3%	5.6pp
1st class: 0.30 (-0.10)	52.1%	58.9%	6.8pp
1st class: 0.50 (+0.10)	55.4%	59.8%	4.4pp
2nd class: 0.15 (-0.10)	53.2%	59.1%	5.9pp
2nd class: 0.35 (+0.10)	54.3%	59.6%	5.3pp
3rd class: 0.00 (-0.10)	54.9%	60.1%	5.2pp
3rd class: 0.20 (+0.10)	52.4%	58.4%	6.0pp

Table 10: Sensitivity to class weight variation. Male survival remains 52–55%, female 58–60%, with sex gap 4–7pp across all specifications. Institutional effect (45–52pp redistribution) substantially exceeds capability bounds (±11pp) in all cases.

Specification	Male Survival	Female Survival	Sex Gap
Baseline	53.7%	59.3%	5.6pp
Family: 0.20 (-0.10)	52.8%	58.7%	5.9pp
Family: 0.40 (+0.10)	54.6%	60.0%	5.4pp
Child: 0.20 (-0.10)	54.2%	58.1%	3.9pp
Child: 0.40 (+0.10)	53.1%	60.6%	7.5pp

Table 11: Sensitivity to family and child weight variation. Qualitative finding persists: institutional effects (45–52pp) far exceed capability variation bounds (±11pp).

F.5 E.3.5 Key Finding

Across all tested specifications (n=12 variations):

- Male survival: 51–57% (range: 6pp)
- Female survival: 58–61% (range: 3pp)
- Institutional effect: 45–52pp
- **Ratio to capability bounds: 4.1–4.7×**

The qualitative finding—that institutional effects substantially exceed plausible individual capability variation—is robust to weight specification. The exact magnitude varies modestly (45–52pp), but remains 4–5× larger than capability bounds under all reasonable parameterizations.

This sensitivity analysis strengthens confidence that our core conclusion (institutional structure dominates individual variation) does not depend on arbitrary weight choices.

F.6 Alternative Capability Bounds

Our capability bounds (±11pp) assume additive effects. Under multiplicative interactions (e.g., physical fitness × decision-making × swimming ability), individual variation could be larger.

To test sensitivity:

- If individual effects reach 20pp (generous multiplicative bound), institutional effects (50pp) still exceed this by 2.5×
- Even under extreme capability assumptions, institutional effects remain substantially larger

This strengthens confidence that institutional mechanisms dominate, though we acknowledge uncertainty about true capability variation in emergency contexts.

G Ethical Considerations

G.1 Respect for Historical Tragedy

This analysis uses data representing 1,502 deaths. We acknowledge the human tragedy and avoid treating the disaster as merely a dataset. Our focus on institutional mechanisms aims to understand how systems can fail, which may inform modern safety and fairness considerations.

G.2 Normative Neutrality

We make no claims about whether 1912 evacuation protocols were morally justified by contemporary standards. Our analysis is descriptive (documenting what happened and how institutional mechanisms operated), not prescriptive (advocating for particular policies).

G.3 Modern Applications

We discuss implications for modern machine learning but do not analyze specific deployed systems. Applying these methods to real-world algorithmic systems requires careful consideration of:

- Stakeholder impacts
- Appropriate transparency
- Regulatory compliance
- Potential unintended consequences of interventions

We provide methodological tools, not policy recommendations.

H Future Research Directions

H.1 Empirical Extensions

1. **Modern applications**: Apply residual analysis to contemporary algorithmic systems with appropriate data (e.g., credit scoring, hiring, college admissions)
2. **Longitudinal settings**: Extend framework to contexts where advantages compound over time (educational trajectories, career progression)
3. **Continuous outcomes**: Test whether insights generalize beyond binary outcomes to continuous variables (income, test scores)
4. **Cross-domain validation**: Analyze multiple historical cases to identify common patterns

H.2 Methodological Development

1. **Formal hypothesis testing**: Develop statistical frameworks for testing "capability model" vs "institutional structure model" hypotheses

2. **Causal identification**: Integrate residual analysis with causal inference methods (instrumental variables, regression discontinuity, natural experiments)

3. **Power analysis**: Provide sample size recommendations for detecting structured residuals of various magnitudes

4. **Risk-aware fairness**: Formalize metrics that incorporate variance and skewness, not just means

H.3 Theoretical Development

1. **Decomposition methods**: Formally separate institutional structure, individual capability, and true randomness components

2. **Multi-level models**: Develop hierarchical frameworks for nested institutional effects

3. **Dynamic models**: Extend to settings where institutional rules evolve over time

References

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231.

British Wreck Commissioner's Inquiry (1912). Report on the loss of the titanic (s.s.). `https://www.gov.uk/government/publications/report-on-the-loss-of-the-titanic-ss-1912`.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2 edition.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, 60(6):581–592.

Kaggle Inc. (2012). Titanic: Machine learning from disaster. `https://www.kaggle.com/c/titanic`.

Lord, W. (1955). *The Night Lives On: The Untold Stories and Secrets Behind the Sinking of the Titanic*. Penguin Books.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4768–4777.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3):289–310.

Zell, E. and Krizan, Z. (2015). Gender differences in domain-specific self-esteem: A meta-analysis. *Review of General Psychology*, 19(4):447–458.