An Integrated Framework for Data and Statistical Analysis in Research

Lamhot Jepri Manarihon Siagian

June 21, 2025

An Integrated Framework for Data and Statistical Analysis in Research

Introduction

The ability to transform raw observations into actionable insights is fundamental to rigorous research. This paper presents an integrated framework covering the hierarchy of data and information, measurement scales, descriptive and bivariate statistical methods, probability and distributions, and sampling techniques. This paper illustrates each concept with empirical examples drawn from a recent survey of research software developers (Eisty, Kanewala, & Carver, 2025). By grounding theory in practice, this work aims to guide researchers in designing robust studies and interpreting quantitative findings effectively.

Part 1: Understanding Data and Measurement

1.1 Data and Information Hierarchy

The Data–Information–Knowledge–Wisdom (DIKW) hierarchy describes how unprocessed observations (data) evolve into contextualized information, interpreted knowledge, and ultimately wise decisions (Ackoff, 1989).

- **Data** are raw symbols or measurements devoid of context. For example, the count of respondents dedicating 1–25% of their time to testing (n = 76) in the research software survey represents data (Eisty, Kanewala, & Carver, 2025).
- **Information** emerges when data are organized meaningfully. Reporting that "59% of participants spend 1–25% of their time on testing" converts counts into interpretable percentages (Eisty et al., 2025).
- **Knowledge** is derived by analyzing patterns: noting that a majority devote minimal time to testing suggests potential under-investment in

quality assurance processes.

- **Wisdom** involves applying knowledge to make informed decisions, such as allocating dedicated testing resources to improve software reliability.

1.2 Variables and Measurement Scales

Stevens (1946) classified variables by their scales of measurement, each permitting specific statistical operations:

Scale	Definition	Example
Nominal	Categories without intrinsic order	Test type (unit, integration, system)
Ordinal	Ranked categories with unequal intervals	Process formality (1 = ad hoc to 7 = fully systematic)
Interval	Equal intervals, no true zero	Temperature in Celsius
Ratio	Equal intervals with a meaningful zero	Number of FTE test engineers (0, 1, 2, …)

Nominal data permit mode calculations only; ordinal allow median and percentiles; interval support addition/subtraction (but not ratios); ratio support full arithmetic operations.

Part 2: Descriptive Statistics and Bivariate Analysis

2.1 Frequency Distribution and Summary Measures

After analyzing the variable **"Percentage of development time spent on testing"** from Eisty et al.'s survey. Table 7 presents the distribution for 128 valid responses:

Time Spent on Testing	Count	Percentage	Midpoint (%)
1–25%	76	59%	12.5
26–50%	42	33%	37.5
51–75%	9	7%	62.5
76–100%	1	1%	87.5

Using midpoints, here is the compute result:

- **Mean:** 24.8%
- **Median:** 12.5% (64th observation falls in 1–25% category)
- **Mode:** 12.5% (most frequent category)
- **Range:** 75% (87.5% – 12.5%)
- **Variance:** 279.3 (percentage^2)

- **Standard Deviation:** 16.7%

These metrics reveal a positively skewed distribution: although the mean suggests a quarter of time spent on testing, most respondents allocate substantially less.

2.2 Bivariate Analysis

2.2.1 Association of Two Qualitative Variables From Table 37, cross-tabulation of dedicated testing FTEs (Yes/No) and documented test requirements (Yes/No):

	Documented: Yes	No	Total
Dedicated FTEs: Yes	21	15	36
Dedicated FTEs: No	32	62	94
Total	53	77	130

Projects with dedicated FTEs have a documentation rate of 58% (21/36) versus 34% (32/94) without dedicated staff. Chi-square test yields $^{2}(1)$ 7.12, p $<$ .01, indicating a significant association.

2.2.2 Correlation of Two Quantitative Variables Using Fisher's Iris dataset (Anderson, 1935), the Pearson correlation between sepal length and petal length (n = 150) is r = .87 (p $<$ .001), demonstrating a strong positive linear relationship.

Part 3: Probability and Distributions

3.1 Rules of Probability

Key principles:

- **Addition rule:** For mutually exclusive A and B, P(A B) = P(A) + P(B).
- **Multiplication rule:** For independent A and B, P(A B) = P(A) · P(B).
- **Conditional probability:** P(A|B) = P(A B)/P(B).
- **Bayes' theorem:**

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}.$$

Example: A test tool with 90% sensitivity, 95% specificity, and 5% base failure rate yields P(failure | signal) 48.6%, illustrating Bayesian updating.

3.2 Random Variables and Distributions

- **Discrete:** Number of test failures in 10 cases follows Binomial(n = 10, p = .05).
- **Continuous:** Execution time ~ Normal(= 120 s, = 10 s), modeling variability.

Part 4: Sampling Techniques

4.1 Random vs. Non-Random Sampling

- **Random sampling** assigns each population member a known chance of selection (e.g., random sample from a developer mailing list).
- **Non-random sampling** (e.g., convenience sampling at conferences) may introduce bias but is pragmatic for hard-to-reach groups.

4.2 Determining Sample Size

For estimating a proportion with margin of error E at confidence level z:

$$n = \frac{z^2\, p\,(1-p)}{E^2}.$$

Example: To estimate within ±5% at 95% confidence (p = .5), n 384 (Cochran, 1977).

Conclusion

This paper offers a unified treatment of foundational data concepts and statistical methods, illustrated by empirical research and classical datasets. By integrating theory with application, researchers can design rigorous studies, conduct appropriate analyses, and derive meaningful insights from quantitative data.

References

- Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis, 16*, 3–9.
- Anderson, E. (1935). The irises of the Gaspé Peninsula. *Bulletin of the American Iris Society, 59*, 2–5.
- Bayes, T., & Price, R. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London, 53*, 370–418.

- Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). John Wiley & Sons.
- Eisty, N. U., Kanewala, U., & Carver, J. C. (2025). Testing research software: An in-depth survey of practices, methods, and tools [Preprint]. *arXiv*. https://arxiv.org/abs/2501.17739
- Ross, S. M. (2014). *A first course in probability* (9th ed.). Pearson.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677–680. https://doi.org/10.1126/science.103.2684.677

Appendix: Python Code

```python
import numpy as np
import pandas as pd
from scipy import stats
from sklearn import datasets

# Part 2.1: Summary statistics for time spent on testing
data = [12.5]*76 + [37.5]*42 + [62.5]*9 + [87.5]*1
mean = np.mean(data)
median = np.median(data)
mode = stats.mode(data).mode[0]
range_ = np.max(data) - np.min(data)
variance = np.var(data)
std_dev = np.std(data)

summary_df = pd.DataFrame({
    'Statistic': ['Mean', 'Median', 'Mode', 'Range', 'Variance', 'Std Dev'],
    'Value': [mean, median, mode, range_, variance, std_dev]
})
print(summary_df)

# Part 2.2.1: Cross-tabulation and chi-square test
ct_df = pd.DataFrame({
    'Dedicated_FTE': ['Yes']*36 + ['No']*94,
    'Documented_Req': ['Yes']*21 + ['No']*15 + ['Yes']*32 + ['No']*62
})
ct = pd.crosstab(ct_df['Dedicated_FTE'], ct_df['Documented_Req'])
chi2, p, dof, exp = stats.chi2_contingency(ct)
print(ct)
print(f"Chi-square: {chi2:.2f}, p-value: {p:.3f}")

# Part 2.2.2: Pearson correlation on Iris dataset
iris = datasets.load_iris()
iris_df = pd.DataFrame(iris.data, columns=iris.feature_names)
```

```
r = iris_df['sepal length (cm)'].corr(iris_df['petal length (cm)'])
print(f"Pearson correlation: {r:.2f}")

# Part 3.2: Probability distributions examples
from scipy.stats import binom, norm
pmf = [binom.pmf(k, 10, 0.05) for k in range(11)]
print("Binomial PMF (n=10, p=0.05):", pmf)
x = np.linspace(80, 160, 100)
pdf = norm.pdf(x, loc=120, scale=10)
print("Normal PDF sample:", pdf[:5])
```