Chapter 4: Students Mostly Harmless Statistics

4.1 Introduction

Modern statistics is often perceived as a technical and mathematically demanding discipline, filled with complex formulas, sophisticated models, and advanced computational tools. While these elements are undoubtedly important, they can sometimes obscure the primary goal of statistical analysis: **learning credible and useful information from data**. The idea of *mostly harmless statistics* emphasizes simplicity, transparency, and careful reasoning over unnecessary complexity.

The phrase "mostly harmless" reflects a philosophy rather than a strict methodology. It suggests that statistical methods should be used in ways that minimize harm—harm in the sense of misleading conclusions, overconfident claims, fragile results, and unjustified causal interpretations. For research students and applied analysts, this perspective is especially valuable because many statistical errors arise not from incorrect calculations but from poor research design, weak assumptions, and misinterpretation of results.

This chapter introduces the principles of mostly harmless statistics, focusing on causal thinking, robustness, design-based approaches, and responsible interpretation. Rather than presenting statistics as a purely mathematical exercise, the chapter frames it as a tool for disciplined reasoning about evidence. The emphasis is on methods that are intuitive, interpretable, and defensible, particularly in applied research settings.

4.2 The Philosophy of Mostly Harmless Statistics

4.2.1 Statistics as a Tool, Not an End

Mostly harmless statistics treats statistical methods as **means to an end**, not ends in themselves. The purpose of analysis is not to demonstrate technical sophistication but to answer substantive research questions clearly and honestly. Overly complex models can give a false sense of precision and may hide strong assumptions that are difficult to justify.

A key principle is that **simpler methods, when well-designed, often outperform complex ones** in terms of credibility and interpretability. This does not imply rejecting advanced techniques, but rather using them judiciously and transparently.

4.2.2 The Cost of Statistical Harm

Statistical harm can occur in several ways:

- Drawing causal conclusions from purely correlational data
- Ignoring violations of model assumptions
- Overfitting models to limited data

- Selective reporting of statistically significant results
- Misinterpreting p-values and confidence intervals

Mostly harmless statistics aims to reduce these risks by prioritizing design, clarity, and robustness.

4.3 Causal Thinking in Statistics

4.3.1 Correlation Is Not Causation

One of the most fundamental lessons in statistics is that correlation does not imply causation. Yet, causal claims are often implicitly or explicitly made in research, policy analysis, and public discourse.

Mostly harmless statistics emphasizes **explicit causal questions**, such as:

- What is the effect of an intervention?
- What would have happened in the absence of treatment?
- How comparable are treated and untreated groups?

By framing analysis around causal questions, researchers are forced to confront assumptions and limitations directly.

4.3.2 Potential Outcomes Framework

The potential outcomes framework provides a conceptual foundation for causal inference. Each unit has:

- A potential outcome under treatment
- A potential outcome under control

The causal effect is defined as the difference between these outcomes. Since only one outcome can be observed for each unit, causal inference becomes a problem of **missing data**. Design-based strategies aim to approximate this missing information.

4.4 Design-Based Approaches to Inference

4.4.1 Why Design Matters More Than Analysis

Mostly harmless statistics prioritizes **research design over statistical adjustment**. A well-designed study with simple analysis is often more credible than a poorly designed study with advanced modeling.

Design-based approaches attempt to create conditions under which treated and control groups are comparable.

4.4.2 Randomized Controlled Experiments

Randomized experiments are the gold standard of causal inference. Random assignment ensures that, on average, treated and control groups are similar in both observed and unobserved characteristics.

Key advantages include:

- Minimal assumptions
- Transparent causal interpretation
- Simple statistical analysis

However, experiments are not always feasible due to ethical, practical, or financial constraints.

4.5 Quasi-Experimental Methods

When randomization is not possible, researchers often rely on quasi-experimental designs.

4.5.1 Difference-in-Differences

Difference-in-differences compares changes over time between treated and control groups. It relies on the assumption of parallel trends.

This method is widely used because it is:

- Intuitive
- Transparent
- Easily visualized

4.5.2 Regression Discontinuity Designs

Regression discontinuity exploits sharp cutoffs in treatment assignment. Units near the cutoff are assumed to be similar, making causal inference plausible.

This approach emphasizes local comparisons and careful graphical analysis.

4.5.3 Instrumental Variables

Instrumental variables isolate exogenous variation in treatment. A valid instrument affects the outcome only through its effect on treatment.

While powerful, instrumental variable methods rely on strong assumptions and should be used cautiously.

4.6 The Role of Regression in Mostly Harmless Statistics

4.6.1 Regression as a Descriptive Tool

In mostly harmless statistics, regression is often viewed as a **descriptive or adjustment tool**, not a magic causal machine. Regression can help control for observed differences, but it cannot fix fundamental design problems.

4.6.2 Overreliance on Controls

Adding many control variables may give the illusion of rigor while increasing sensitivity to model assumptions. Mostly harmless approaches encourage:

- Parsimonious models
- Justified control selection
- Sensitivity analysis

4.7 Robustness and Sensitivity Analysis

4.7.1 Why Robustness Matters

Robustness checks assess whether results depend heavily on specific modeling choices. Fragile results are a warning sign of potential statistical harm.

4.7.2 Common Robustness Checks

Examples include:

- Alternative model specifications
- Subsample analysis
- Different functional forms
- Placebo tests

Robust findings inspire greater confidence.

4.8 Statistical Significance Reconsidered

4.8.1 Limitations of p-Values

Mostly harmless statistics views p-values as limited tools rather than decisive evidence. Overemphasis on arbitrary significance thresholds can distort research incentives and interpretation.

4.8.2 Practical vs. Statistical Significance

Researchers are encouraged to focus on:

- Effect sizes
- Confidence intervals
- Substantive relevance

Statistical significance alone does not imply importance.

4.9 Visualization as a Core Analytical Tool

Graphs and visualizations play a central role in mostly harmless statistics. Visual analysis can:

- Reveal patterns hidden in tables
- Diagnose design assumptions
- Communicate results clearly

Common tools include:

- Scatterplots
- Time-series plots
- Regression discontinuity graphs

Visualization promotes transparency and intuition.

4.10 Transparency and Reproducibility

4.10.1 Open Data and Code

Mostly harmless statistics supports reproducible research practices, including:

- Sharing data
- Publishing code
- Clear documentation

Transparency reduces errors and increases trust.

4.10.2 Pre-Analysis Plans

Pre-analysis plans help prevent data mining and selective reporting by committing researchers to analysis strategies in advance.

4.11 Ethical Dimensions of Statistical Practice

Statistical harm can have real-world consequences, especially in policy, medicine, and social research. Ethical statistical practice requires:

- Honest reporting
- Acknowledgment of limitations
- Avoidance of overclaiming

Mostly harmless statistics aligns with ethical research principles.

4.12 Applications Across Disciplines

The mostly harmless approach has influenced:

- Economics
- Public policy
- Education research
- Epidemiology
- Social sciences

Its emphasis on design and credibility makes it broadly applicable.

4.13 Common Misuses of Statistics

Common errors include:

- Treating regression coefficients as causal effects without justification
- Ignoring selection bias
- Overinterpreting noisy estimates
- Failing to check assumptions

Mostly harmless statistics encourages skepticism and humility.

4.14 Learning Statistics the Mostly Harmless Way

For research students, learning statistics through this lens involves:

- Asking clear causal questions
- Prioritizing design

- Using simple, transparent methods
- Interpreting results cautiously

This approach builds long-term analytical competence.

4.15 Case Study: Policy Evaluation

Consider an evaluation of a job training program. Rather than relying solely on regression controls, a mostly harmless approach might use random assignment or a difference-in-differences design, supported by graphical analysis and robustness checks. This produces more credible conclusions.

4.16 Summary

Mostly harmless statistics emphasizes careful reasoning, credible design, and responsible interpretation. By focusing on transparency, robustness, and causal thinking, researchers can reduce statistical harm and produce findings that are both informative and trustworthy. This philosophy does not reject technical methods but places them within a disciplined framework that prioritizes understanding over complexity.

References

1. Angrist, J. D., & Pischke, J. S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
2. Angrist, J. D., & Pischke, J. S. (2015). *Mastering 'Metrics*. Princeton University Press.
3. Hernán, M. A., & Robins, J. M. (2020). *Causal Inference: What If*. Chapman & Hall/CRC.
4. Gelman, A., et al. (2014). *Bayesian Data Analysis*. CRC Press.
5. Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
6. Freedman, D. A. (2010). *Statistical Models and Causal Inference*. Cambridge University Press.
7. Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values. *The American Statistician*, 70(2), 129–133.