SMPE

January 21, 2025

1 Design of Experiments

The lecture on the "Design of Experiments" covered foundational principles and methodologies for designing and analyzing experiments in a systematic and efficient manner. Below, we summarize the main concepts and methods discussed:

1.1 Key Concepts

Two essential principles underpin successful experimental design:

- **Replication**: Increases the reliability of results by performing experiments multiple times.
- **Randomization**: Reduces bias by ensuring experimental conditions are assigned randomly.

Additional considerations include avoiding pseudo-replication and distinguishing between experimental and observational data. The lecture emphasized that even the most advanced statistical analysis cannot correct for poorly designed experiments.

1.2 Defining the Problem and Variables

Successful experiments begin with clearly defining the problem:

- Determine the system and phenomena to study.
- Decide on the type of study (descriptive, exploratory, predictive, hypothesis testing, etc.).

The system is typically modeled as a black box with controllable inputs $(x_1, \dots, x_p)$, uncontrollable inputs $(z_1, \dots, z_q)$, and outputs (y). Identifying relevant response variables (e.g., makespan, energy usage, convergence time) and controllable factors (e.g., algorithm heuristics, platform size) is critical.

1.3 Experimental Design Approaches

Several experimental designs were introduced to balance the trade-off between accuracy and feasibility:

- **2-Level Factorial Designs**: Test every combination of high (+1) and low (-1) values for p factors, enabling detection of interactions and estimation of main effects.
- **Fractional Factorial Designs**: Reduce the number of experiments for large p by considering only a subset of combinations while maintaining balance and spread.
- **Screening Designs**: Identify primary factors with significant effects using methods like Plackett-Burman designs.
- **General Factorial Designs**: Extend factorial designs to factors with more than two levels, though analysis becomes complex with increasing levels.

1.4 Sequential Approach and Parsimony

The lecture advocated for a sequential approach to experimentation:

1. Begin with exploratory designs to identify key factors.
2. Use parsimonious models, retaining only significant factors and interactions.
3. Add experiments iteratively, focusing on areas of high variability.

The principle of parsimony emphasizes simplicity: models should be as simple as possible while adequately explaining the data.

1.5 Analysis and Optimization

Analyzing experimental data often involves:

- **Regression Models**: Fit linear or polynomial models to describe the relationship between factors and response variables.
- **Analysis of Variance (ANOVA)**: Discriminate real effects from noise, assuming normality and homoscedasticity.
- **Optimal Designs**: Use D-optimal or I-optimal designs to minimize variance in model parameter estimates.

1.6 Techniques for Analysis

1.6.1 Analysis of Variance (ANOVA)

ANOVA is a statistical technique to compare means among multiple groups and determine if at least one mean differs significantly. The main components include:

- **Model**: $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, where:
  - y_{ij}: Response variable for group i, observation j.
  - μ: Overall mean.
  - α_i: Effect of group i.
  - ϵ_{ij}: Random error, $\epsilon_{ij} \sim N(0, \sigma^2)$.
- **F-statistic**: $F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$, where:
  - MS_{between}: Mean square for group differences.
  - MS_{within}: Mean square for residual errors.
- **Hypothesis**: H_0: All group means are equal; H_a: At least one group mean differs.

1.6.2 Bayesian Methods

Bayesian approaches incorporate prior knowledge and update beliefs based on observed data. Key elements include:

- **Bayes' Theorem**: $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$, where:
  - $P(\theta|D)$: Posterior probability of parameters θ given data D.
  - $P(D|\theta)$: Likelihood of data given θ.
  - $P(\theta)$: Prior probability of θ.
  - $P(D)$: Marginal probability of data.
- **Advantages**: Flexibility, ability to incorporate uncertainty and prior knowledge.
- **Tools**: Markov Chain Monte Carlo (MCMC) for posterior estimation.

1.6.3 Regression Analysis

Regression analysis explores relationships between dependent and independent variables:

- **Simple Linear Regression**: $y = \beta_0 + \beta_1 x + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$.
- **Multiple Regression**: $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$.
- **Assessment**: R^2 for goodness of fit, F-test for model significance.

1.6.4 Other Techniques

- **Taguchi Methods**: Robust designs to minimize variability.
- **Response Surface Methodology (RSM)**: Explore relationships between factors and responses, identify optimal conditions.
- **Non-Parametric Methods**: Useful when data do not meet parametric assumptions.

1.7 Conclusion

The lecture concluded with the importance of a thoughtful, sequential approach to experimentation. Key takeaways include:

- Proper design and modeling are paramount.
- Use statistical tools judiciously, focusing on principles like replication, randomization, and parsimony.
- Employ suitable designs for the problem at hand, iterating as necessary.

Analysis for Scientific Methodology and Performance Evaluation Exam

Good Graphics (˜0h20)

Q1.1: Figure Analysis The figure claims to represent gender parity in primary education across various countries. The text provides a Gender Parity Index (GPI) threshold of 0.97 to 1.03 for gender parity. However, it fails to clarify how this threshold applies in the depicted graph, which might make certain disparities less apparent. For example:

- The GPI thresholds may not be visually evident, misleading the reader to conclude gender parity for all countries.
- The data presentation might oversimplify nuanced disparities, especially for countries outside the parity range.

Q1.2: Improved Representation A better representation could involve:

- A bar chart indicating GPI values for each country, with regions below 0.97 and above 1.03 highlighted in red or blue.
- Confidence intervals for each data point to indicate variability.

Sketch: Include labeled bars categorized as: *Bias against Girls* (GPI < 0.97), *Parity Achieved* ($0.97 \leq$ GPI ≤ 1.03), and *Bias against Boys* (GPI > 1.03).

Open Science and Reproducible Research (˜0h40)

Q2.1: Techniques and Precautions To ensure research reproducibility, the following techniques should be implemented:

Version Control (Git): Prevents confusion caused by multiple dataset or code versions.

- **Why:** Avoids conflicts and duplication.
- **How:** Track changes in research artifacts.
- **When:** From project inception.

Data Documentation: Use metadata and clear data dictionaries.

- **Why:** Clarifies dataset structure and prevents misinterpretation.
- **How:** Include headers and README files.
- **When:** During data collection or curation.

Pre-registration of Studies: Specifies objectives and hypotheses beforehand.

- **Why:** Minimizes post hoc bias.
- **How:** Utilize pre-registration platforms.
- **When:** Before conducting the experiments.

(*... Include other techniques as required*)

Data Analysis (˜0h50)

Q3.1: Opinion on Regression Analysis The analysis suffers from the following flaws:

- The chosen polynomial model may not fit well for all workload ranges.
- Insufficient diagnostics: Residuals display patterns indicating potential model misfits.

Q3.2: Recommendations

- Perform exploratory data analysis (EDA) before regression.
- Consider alternative models (e.g., mixed effects).
- Validate models using cross-validation.

Q3.3: Efficient Experimental Design

- **Assumption:** Response time increases monotonically with load.
- **Design:** Utilize adaptive sampling near thresholds to gather more granular data.
- **Framework:** Use hypothesis testing and confidence intervals to delineate thresholds.

Article Analysis (˜1h00)

Q4.1: Summary of the Article The article highlights three challenges in policing strategy studies:

1. **Measurement Issues:** Differences in crime definitions skew datasets.
2. **Correlation vs. Causation:** Statistical analysis struggles to isolate cause-effect relationships.
3. **Intervention Prediction:** Outcomes vary with geographic and societal contexts.

Q4.2: Reproducibility Challenges: Disparate datasets and statistical pitfalls make replication difficult.
Q4.3: Disagreement (Your identified statements will be critiqued here)
Q4.4: AI Ethical Analysis

- **Problem:** Reducing prediction uncertainty in crime prevention.
- **Helpful:** Scales analysis and aids precision.
- **Cons:** Risks of bias propagation.
- **ADEME Scenario:** Green technologies or regional cooperation.

Keywords and Concepts for Exam Preparation

Core Concepts in SMPE

1. Reproducible Research:

- Definition and importance in scientific research.
- Best practices: Use of version control (e.g., Git), data documentation, and pre-registration.
- Challenges: Lack of standardization and missing metadata.

2. Performance Evaluation:

- Understanding experimental design for evaluating algorithms.
- Key metrics: precision, recall, latency, throughput.
- Use of benchmarking and simulation tools.

3. Data Visualization:

- Principles: clarity, avoiding misleading representations.
- Tools: ggplot2, matplotlib.
- Techniques: histograms, scatter plots, bar charts.

Special Topics

1. Open Science:

- The ethos of open-access journals.
- Licensing issues in data sharing (e.g., GPL, MIT licenses).

2. Ethical Concerns in Data Analysis:

- Addressing bias in AI and statistical models.
- Data privacy and protection (GDPR context).

Technical Skills to Master

1. Programming for Analysis:

- R: tidyr, dplyr, ggplot2.
- Python: pandas, NumPy, matplotlib.

2. Statistical Techniques:

- Linear regression (modeling and assumptions).
- Hypothesis testing (p-values, t-tests).

3. Experiment Design:

- Concept of control groups.
- Adaptive sampling techniques.

2 Design of Experiments

The lecture on the "Design of Experiments" covered foundational principles and methodologies for designing and analyzing experiments in a systematic and efficient manner. Below, we summarize the main concepts and methods discussed:

2.1 Key Concepts

Two essential principles underpin successful experimental design:

- **Replication**: Increases the reliability of results by performing experiments multiple times.
- **Randomization**: Reduces bias by ensuring experimental conditions are assigned randomly.

Additional considerations include avoiding pseudo-replication and distinguishing between experimental and observational data. The lecture emphasized that even the most advanced statistical analysis cannot correct for poorly designed experiments.

2.2 Defining the Problem and Variables

Successful experiments begin with clearly defining the problem:

- Determine the system and phenomena to study.
- Decide on the type of study (descriptive, exploratory, predictive, hypothesis testing, etc.).

The system is typically modeled as a black box with controllable inputs $(x_1, \ldots, x_p)$, uncontrollable inputs $(z_1, \ldots, z_q)$, and outputs (y). Identifying relevant response variables (e.g., makespan, energy usage, convergence time) and controllable factors (e.g., algorithm heuristics, platform size) is critical.

2.3 Experimental Design Approaches

Several experimental designs were introduced to balance the trade-off between accuracy and feasibility:

- **2-Level Factorial Designs**: Test every combination of high (+1) and low (-1) values for p factors, enabling detection of interactions and estimation of main effects.
- **Fractional Factorial Designs**: Reduce the number of experiments for large p by considering only a subset of combinations while maintaining balance and spread.
- **Screening Designs**: Identify primary factors with significant effects using methods like Plackett-Burman designs.
- **General Factorial Designs**: Extend factorial designs to factors with more than two levels, though analysis becomes complex with increasing levels.

2.4 Sequential Approach and Parsimony

The lecture advocated for a sequential approach to experimentation:

1. Begin with exploratory designs to identify key factors.
2. Use parsimonious models, retaining only significant factors and interactions.
3. Add experiments iteratively, focusing on areas of high variability.

The principle of parsimony emphasizes simplicity: models should be as simple as possible while adequately explaining the data.

2.5 Analysis and Optimization

Analyzing experimental data often involves:

- **Regression Models**: Fit linear or polynomial models to describe the relationship between factors and response variables.
- **Analysis of Variance (ANOVA)**: Discriminate real effects from noise, assuming normality and homoscedasticity.
- **Optimal Designs**: Use D-optimal or I-optimal designs to minimize variance in model parameter estimates.

2.6 Conclusion

The lecture concluded with the importance of a thoughtful, sequential approach to experimentation. Key takeaways include:

- Proper design and modeling are paramount.
- Use statistical tools judiciously, focusing on principles like replication, randomization, and parsimony.
- Employ suitable designs for the problem at hand, iterating as necessary.

3 Causality, Dependency, Correlation, and Designed Experiments

3.1 Overview

This lecture explores the distinctions and interplay between correlation, causation, dependency, and the methodologies of designed experiments. It emphasizes the importance of critical thinking in interpreting data, avoiding pitfalls such as spurious correlations, and utilizing sound experimental design principles.

3.2 Correlation and Dependency

The relationship between two variables, X and Y, can be quantified using the correlation coefficient:

$$\operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y},$$

where $\operatorname{cov}(X,Y)$ is the covariance and σ_X, σ_Y are the standard deviations of X and Y. Key properties include:

- $\operatorname{corr}(X,Y)$ lies in $[-1,1]$.
- Perfect linear relationships yield $\operatorname{corr}(X,Y) = \pm 1$.
- Independence implies $\operatorname{corr}(X,Y) = 0$, though the converse is not necessarily true.

While correlation suggests a relationship, it does not imply causation. For example, spurious correlations often arise due to shared dependence on a third variable or random coincidence.

3.3 Illustrative Examples

- **Global Temperature and Number of Pirates:** Demonstrates how two unrelated variables can appear correlated due to a shared temporal trend.
- **Bee Colonies and Divorce Rates:** Highlights the absurd conclusions that can result from misinterpreting spurious correlations.

3.4 Causation and Experimental Design

To establish causation, well-designed experiments are critical. Observational data, while valuable, often cannot definitively test hypotheses due to confounding variables and biases. Experimental data, characterized by replication and randomization, allows for more reliable inferences. Key principles include:

- **Replication:** Ensures reliability of results by repeating experiments under similar conditions.
- **Randomization:** Minimizes bias by randomly assigning subjects or conditions.
- **Avoiding Pseudo-replication:** Ensures measurements represent truly independent observations.

3.5 Case Study: The Linux Users Study

An example discussed was a satirical study comparing the physical attributes of Linux users with Windows users. The study suffered from numerous biases:

- Lack of proper random sampling.
- Ignoring outliers and variability.
- Misinterpretation of averages and failure to consider confounding factors.

The analysis humorously illustrated the dangers of drawing conclusions from poorly designed experiments.

3.6 Conclusion

This lecture underscores the necessity of blending observation, theory, and experimentation to derive meaningful insights. Properly designed experiments, supported by clear hypotheses and robust statistical analysis, are indispensable for distinguishing correlation from causation and for advancing scientific understanding.