01 Introduction And Motivation

Problem Statement

The chapter addresses the challenge of accurately assessing the potential impact of new ideas in online environments, particularly through the use of controlled experiments (A/B tests).

Motivating Example

In 2012, a simple suggestion to alter the ad headline display on Bing led to a significant increase in revenue, demonstrating the unpredictable value of minor changes. This change, initially underestimated and deprioritized, eventually increased Bing's revenue by 12%, translating to over $100M annually in the US alone. This outcome underscores the importance of empirical testing over intuition or expert opinions in evaluating new ideas.

Solutions and Methods

- **Controlled Experiments:** Utilize A/B tests to empirically evaluate the impact of changes in a controlled manner, ensuring that users are randomly assigned to either the control or treatment group to accurately measure the effect of modifications.
  - **Randomization:** Ensure that the assignment of users to different variants is statistically sound to maintain the integrity of the experiment.
  - **Metrics:** Define clear metrics (Overall Evaluation Criterion - OEC) that reflect the goals of the experiment, balancing revenue generation with user experience.
- **Iterative Testing:** Implement changes incrementally and test continuously to identify and build on successful modifications.
- **Scalability of Experiments:** Leverage platforms that support large-scale testing environments to handle extensive user bases and multiple test variants.
- **Long-term and Multi-faceted Evaluation:** Replicate successful experiments over extended periods and across various metrics to confirm findings and assess broader impacts.
- **Infrastructure for Experimentation:** Develop and maintain robust systems (like Microsoft's ExP) that can support the complex requirements of controlled online experiments.

02 Running And Analyzing Experiments

Problem Statement

The chapter addresses the challenge of designing, running, and analyzing controlled experiments to evaluate the impact of UI changes on user behavior and revenue, specifically through the example of adding a coupon code field to a checkout page.

Motivating Example

The motivating example involves a fictional online commerce site considering the introduction of a coupon code in the checkout process to potentially increase sales. This decision is complicated by conflicting external data suggesting both potential revenue loss and gains from similar actions by other companies. The experiment aims to assess the impact of this UI change on the checkout process and overall revenue, using a simple A/B test setup with two different UI implementations of the coupon code field.

Solutions

- **Experiment Design:**
  - Define the hypothesis: Adding a coupon code field will degrade revenue-per-user.
  - Select metrics: Revenue-per-user, focusing on users who start the purchase process.

  - Determine experiment size and duration to ensure statistical power and capture day-of-week effects.

- **Statistical Significance:**
  - Calculate baseline mean and standard error to understand metric variability.
  - Use p-value and confidence intervals to test the null hypothesis that the treatment and control groups have the same mean revenue-per-user.

- **Practical Significance:**
  - Establish a practical significance boundary to determine if observed differences are meaningful from a business perspective.
  - Consider the business impact of a 1% or larger increase in revenue-per-user as significant.

- **Running the Experiment:**
  - Implement necessary infrastructure for variant assignment and data collection.
  - Ensure the experiment includes adequate user coverage and duration to account for variability in user behavior and external factors.

- **Data Analysis and Interpretation:**
  - Perform sanity checks using invariant metrics to validate the experiment's integrity.
  - Analyze the results to determine if the changes are statistically and practically significant.

- **Decision Making:**
  - Use the results to make informed decisions about whether to implement the UI change across all users.
  - Consider both statistical and practical significance in the context of business goals and costs.

03 Twymans Law And Experimentation Trustworthiness

Problem Statement

The chapter delves into the challenges of ensuring the trustworthiness of experimental results in statistics and machine learning, particularly focusing on the implications of Twyman's Law. This law suggests that any statistical result that appears interesting or different is likely to be incorrect, prompting a deeper investigation into the reliability of experimental findings.

Motivating Example

A significant improvement in a key metric from an experiment might lead to premature celebration or, conversely, a negative result might be dismissed too quickly. For instance, an A/B test showing a drastic increase in user engagement could be influenced by errors in data collection, logging issues, or statistical anomalies rather than a genuine effect of the tested changes.

Solutions and Best Practices

- **Statistical Power Analysis:** Ensure experiments are sufficiently powered to detect practical effects, avoiding false negatives due to underpowered studies.
- **P-value Interpretation:** Correct common misconceptions about p-values, such as misunderstanding them as the probability that the null hypothesis is true.
- **Sequential Testing:** Use methods like always valid p-values or Bayesian frameworks to handle continuous monitoring of p-values without inflating the type I error rate.
- **Handling Multiple Hypotheses:** Apply corrections for multiple comparisons to avoid false discoveries when multiple hypothesis tests are conducted.

- **Confidence Intervals:** Use confidence intervals to assess the range of plausible values for an effect size, understanding their relationship with hypothesis tests.
- **Internal Validity Checks:** Address potential violations of the Stable Unit Treatment Value Assumption (SUTVA) and other threats to internal validity such as survivorship bias and sample ratio mismatch (SRM).
- **External Validity and Generalization:** Evaluate how well experimental results generalize across different settings and time periods, considering factors like novelty effects and the stability of treatment effects over time.
- **Segmentation and Simpson's Paradox:** Be cautious of segmenting results which can lead to misleading conclusions if not handled properly, and be aware of Simpson's paradox in aggregated data.

04 Experimentation Platform And Culture

Problem Statement

The chapter addresses the challenge of building a robust and trustworthy experimentation platform that facilitates controlled experiments to evaluate ideas and make data-informed decisions. It emphasizes the importance of an experimentation culture within organizations to accelerate innovation and learning.

Motivating Example

The chapter begins with a quote from Mike Moran, "If you have to kiss a lot of frogs to find a prince, find more frogs and kiss them faster and faster," illustrating the need for rapid experimentation to find successful innovations. It discusses the journey of organizations as they evolve their experimentation capabilities through various maturity phases, highlighting the critical role of leadership and technical infrastructure in fostering a culture that supports continuous experimentation and learning.

Solutions and Methods

- **Experimentation Maturity Models:**
  - *Crawl Phase:* Focus on building foundational capabilities like instrumentation and basic data science skills.
  - *Walk Phase:* Develop standard metrics and improve trust in the experimentation process through validation techniques like A/A tests.
  - *Run Phase:* Scale up the experimentation process, establish comprehensive metrics, and use experiments to evaluate most new features.
  - *Fly Phase:* Integrate A/B testing as a norm for every change, automate processes, and build institutional memory to learn from past experiments.
- **Leadership and Process:**
  - Engage executives in setting metric-based goals rather than feature-based goals.
  - Empower teams to innovate within organizational guardrails.
  - Establish a culture of intellectual integrity where learning from experiments is valued over immediate outcomes.
- **Technical Infrastructure:**
  - Build or buy decision-making based on the scale of experimentation and specific needs.
  - Develop robust systems for experiment definition, setup, deployment, and analysis.
  - Ensure high data quality and integrate systems for real-time results and automated detection of issues.
- **Educational Processes and Cultural Norms:**

  - Implement training programs and just-in-time learning processes to improve the design and analysis of experiments.
  - Use regular review meetings to discuss and learn from experiment outcomes.

05 Speed Matters An End To End Case Study

Problem Statement

The chapter addresses the critical impact of web performance on user satisfaction, revenue, and overall operational efficiency. It emphasizes the necessity of understanding and improving latency to enhance key business metrics.

Motivating Example

The chapter begins with a real-world scenario involving companies like Bing and Amazon, where even a 100 millisecond improvement in load time could lead to significant revenue increases. For instance, a 100 msec speedup at Bing improved revenue by 0.6%, and a similar slowdown at Amazon decreased sales by 1%. These examples underscore the direct correlation between speed and financial performance.

Solutions and Techniques

- **Slowdown Experiments:** Controlled experiments where the product is intentionally slowed to measure the impact on key metrics like revenue and user satisfaction.
  - *Assumption:* Performance improvements or degradations have a linear impact on revenue and user metrics around the current performance level.
  - *Methodology:* Comparing metrics at different slowdown intervals (e.g., 100 msec and 250 msec) to validate the linearity assumption.
- **Measurement Techniques:**
  - *Page Load Time (PLT):* Approximated by measuring the time difference between the initial request and the onload event beacon reaching the server.
  - *Navigation Timing API:* Utilizing browser capabilities to provide more accurate timing metrics across different stages of page loading.
- **Experiment Design Considerations:**
  - *Placement of Slowdown:* Deciding the optimal point to introduce a delay, such as delaying the server response after computing the URL-dependent HTML (Chunk2).
  - *Delay Duration:* Balancing between a delay long enough to measure impacts accurately and short enough to minimize user inconvenience.
- **Perceived Performance Metrics:**
  - *Above the Fold Time (AFT):* Time until the visible part of the page is displayed.
  - *Speed Index:* A measure of how quickly the contents of a page are visibly populated.

Impact of Different Page Elements

The impact of performance improvements varies across different elements of a page. For example, elements that load "below the fold" or in less critical parts of the interface (like the right pane in Bing) have a lesser impact on key metrics compared to elements that users interact with immediately.

06 Organizational Metrics

Problem Statement

The chapter addresses the challenge of defining, measuring, and utilizing organizational metrics effectively to track and improve organizational performance.

Motivating Example

Consider a company like Google, whose mission is to organize the world's information. The goal metrics for Google might directly relate to user engagement and information accessibility. However, defining these metrics is complex because they must encapsulate broad organizational goals like improving user satisfaction and system efficiency, while also being resistant to manipulation and straightforward enough for stakeholders to understand and act upon.

Solutions and Techniques

- **Defining Goal and Driver Metrics:**
  - Goal metrics should be simple and stable, reflecting the ultimate success of the organization.
  - Driver metrics need to be aligned with goal metrics, actionable, sensitive, and resistant to gaming.
- **Frameworks for Driver Metrics:**
  - HEART framework (Happiness, Engagement, Adoption, Retention, Task Success).
  - PIRATE framework (Acquisition, Activation, Retention, Referral, Revenue).
- **Guardrail Metrics:**
  - Protect business operations and assumptions, such as maintaining user engagement levels while increasing user registrations.
- **Metric Formulation Techniques:**
  - Use qualitative concepts to define concrete, quantifiable metrics.
  - Validate driver metrics through experiments to confirm their impact on goal metrics.
  - Incorporate quality dimensions into metrics to ensure they drive desired outcomes.
- **Evaluating and Evolving Metrics:**
  - Continuously validate and refine metrics to adapt to changes in business focus or external environment.
  - Use a combination of data sources and experimental designs to validate the causal relationships of metrics.

Additional Considerations

- Metrics should be iteratively refined as the organization's understanding evolves and as external conditions change.
- Discussions on metrics alignment at all levels of the organization are crucial for ensuring that everyone is focused on the same objectives.

07 Metrics For Experimentation And The Overall Evaluation Criterion

Problem Statement

The chapter addresses the challenge of selecting and refining metrics that are appropriate for evaluating online controlled experiments, particularly in the context of aligning these metrics with long-term business goals.

Motivating Example

Consider an online retailer experimenting with different homepage designs to increase user engagement. The retailer must choose metrics that not only reflect immediate outcomes like click-through rates but also align with broader objectives such as customer satisfaction and retention. This scenario illustrates the need for metrics that are sensitive, attributable, and timely, ensuring they provide meaningful feedback on the experimental variations being tested.

Solutions

- **Selection of Appropriate Metrics:**
  - Metrics must be *measurable* and *attributable* to specific experimental conditions.
  - Sensitivity and timeliness are crucial to detect significant changes and make prompt decisions.
- **Surrogate Metrics:**
  - Use surrogate metrics for long-term outcomes that are not immediately observable, such as customer lifetime value or renewal rates.
- **Granular and Diagnostic Metrics:**
  - Employ more granular metrics for detailed insights into specific features or user interactions.
  - Include diagnostic metrics to investigate specific issues or anomalies during experiments.
- **Combining Metrics into an Overall Evaluation Criterion (OEC):**
  - Integrate multiple key metrics into a single composite metric that reflects broader business objectives.
  - Normalize metrics to a common scale and assign weights reflecting their relative importance to the business goals.
- **Iterative Refinement:**
  - Continuously refine the OEC based on experimental outcomes and evolving business strategies.
  - Ensure the OEC remains aligned with long-term objectives and is not susceptible to gaming or manipulation.

Examples of OEC Implementation

- **Amazon's Email Campaigns:**
  - Initially used revenue from click-throughs as the OEC, leading to an increase in email volume and user complaints.
  - Revised the OEC to account for unsubscribe rates, balancing short-term revenue against long-term customer engagement.
- **Bing's Search Engine:**
  - Faced with the challenge of improving query share and revenue without compromising search quality.
  - Adjusted the OEC to focus on user satisfaction and task completion, rather than just increasing query volume.

08 Institutional Memory And Meta Analysis

Problem Statement

The chapter addresses the critical role of institutional memory in leveraging historical experiment data to enhance future innovations, improve experimentation culture, and refine decision-making processes in organizations.

Motivating Example

Institutional memory at companies like Bing and LinkedIn has enabled these organizations to accumulate comprehensive data from numerous experiments, which in turn has significantly contributed to their strategic goals. For instance, Bing Ads demonstrated how revenue gains between 2013 and 2015 were directly linked to incremental improvements from hundreds of experiments. This historical data provides a robust foundation for meta-analysis, fostering a culture of experimentation and continuous improvement.

Solutions and Methods

- **Experiment Culture Development:**
  - Analyzing past experiments to reinforce the importance of experimentation.
  - Sharing impactful or surprising experiment results to maintain an engaged and informed experimental culture.
  - Regular reporting on experiments affecting key metrics to enhance transparency and accountability.
- **Adherence to Best Practices:**
  - Meta-analysis to identify and promote best experimental practices across teams.
  - Automation of experiment ramp schedules to ensure consistency and efficiency in testing procedures.
- **Inspiring Future Innovations:**
  - Utilizing historical data to prevent repeating past mistakes and to spark effective new initiatives.
  - Analyzing patterns from past experiments to predict the impact of similar future experiments.
- **Enhanced Metric Utilization:**
  - Studying metric sensitivity and related metrics to refine metric selection and usage.
  - Developing probabilistic priors for Bayesian approaches based on historical data to improve experiment evaluation.
- **Empirical Research Facilitation:**
  - Providing a rich dataset for researchers to validate theories and explore new hypotheses through meta-analysis.
  - Investigating statistical biases and proposing correction methods to refine data interpretation and application.

09 Ethics In Controlled Experiments

Problem Statement

The chapter addresses the ethical considerations necessary in designing and conducting controlled experiments, particularly in fields impacting human subjects directly such as biomedical, behavioral research, and technology.

Motivating Example

The chapter discusses two poignant examples: Facebook and Cornell's study on emotional contagion and OKCupid's matchmaking algorithm deception. These studies manipulated user experiences to observe behavioral changes, raising significant ethical concerns about consent, risk, and the manipulation of information.

Solutions and Considerations

- **Respect for Persons:**
  - Ensure transparency and truthfulness in experiments.
  - Obtain informed consent when possible, respecting participant autonomy.
- **Beneficence:**
  - Minimize risks and maximize benefits for participants.
  - Evaluate the necessity and proportionality of the risks involved.

- **Justice:**
  - Ensure fair distribution of risks and benefits.
  - Avoid exploitation of participants in any form.
- **Risk Assessment:**
  - Identify and mitigate potential physical, psychological, and social risks.
  - Consider equipoise to ensure genuine uncertainty in treatment effects.
- **Benefit Analysis:**
  - Clearly define and communicate the potential benefits to participants and society.
  - Consider long-term benefits and indirect benefits such as improved product sustainability.
- **Data Collection and Privacy:**
  - Adhere to privacy by design principles.
  - Ensure data collection is transparent, necessary, and secure.
  - Regularly audit data usage and access.
- **Cultural and Processual Integrity:**
  - Foster an ethical culture that promotes regular ethical reviews and education.
  - Implement robust processes akin to Institutional Review Boards (IRBs) to oversee experiments.

10 Complementary Techniques

Problem Statement

The chapter addresses the need for complementary techniques in the context of A/B testing and experimentation platforms, emphasizing the importance of generating ideas, validating metrics, and establishing evidence when controlled experiments are not possible or sufficient.

Motivating Example

Consider the challenge of finding a reliable proxy metric for user satisfaction, which is inherently difficult to measure directly. By conducting a survey to gather self-reported user satisfaction data and analyzing corresponding large-scale observational metrics, one can identify potential proxies. These proxies can then be validated through controlled experiments to ensure their reliability and effectiveness in representing user satisfaction.

Solutions and Techniques

- **Logs-based Analysis**
  - Helps build intuition about user behavior and system performance by analyzing metrics like session-per-user and click-through rates.
  - Facilitates the characterization of potential metrics by understanding their variance and correlation with existing metrics.
  - Generates ideas for A/B experiments by exploring underlying data, identifying significant drop-offs, and observing user interactions.
- **Human Evaluation**
  - Involves paying human judges to complete tasks and using their responses for subsequent analysis, which is common in search and recommendation systems.
  - Provides calibrated labeled data that complements data gathered from real users, useful for evaluating A/B experiments and debugging.

- **User Experience Research (UER)**
  - Conducts in-depth studies with a small number of users to generate ideas and insights through direct observation and interaction.
  - Utilizes special equipment like eye-tracking and diary studies to gather data that cannot be captured through regular instrumentation.
- **Focus Groups**
  - Engages users in guided discussions to explore a range of topics, useful for early-stage idea validation and understanding emotional reactions.
  - Helps in gathering feedback on ill-formed hypotheses and furthering the design process.
- **Surveys**
  - Employs structured questionnaires to gather data on aspects not observable via instrumentation, such as offline behaviors or user satisfaction post-purchase.
  - Challenges include designing unbiased questions and dealing with self-reported, potentially untruthful answers.
- **External Data**
  - Utilizes data collected by external parties to validate business metrics and generate ideas for measurable proxies.
  - Includes data from academic research, industry benchmarks, and competitive studies to establish general trends and correlations.

11 Observational Causal Studies

Problem Statement

The chapter addresses the challenge of establishing causality in scenarios where randomized controlled experiments are not feasible. This includes situations where the causal action is not under the control of the organization, there are too few units for a statistically significant experiment, or ethical and logistical constraints prevent experimental manipulation.

Motivating Example

Consider the scenario where a company wants to understand the impact on product engagement when users switch from an iPhone to a Samsung phone. This is a typical example where the causal action (phone switching) is not under the direct control of the organization, and thus, a randomized controlled trial is not possible. The chapter uses this example to explore methods for estimating causal effects using observational data.

Solutions and Methods

- **Interrupted Time Series (ITS)**:
  - Uses multiple pre- and post-intervention measurements to estimate the counterfactual.
  - Example: Estimating the effect of police helicopter surveillance on home burglaries by implementing and withdrawing surveillance multiple times.
- **Interleaved Experiments**:
  - Commonly used in evaluating changes to ranking algorithms by interspersing results from two algorithms and measuring engagement metrics like click-through rates.
- **Regression Discontinuity Design (RDD)**:
  - Applicable when there is a clear threshold defining treatment and control groups.
  - Example: Studying the impact of scholarships awarded based on a grade cutoff.

- **Instrumental Variables (IV)**:
  - Attempts to approximate random assignment using naturally occurring 'instruments'.
  - Example: Using the Vietnam War draft lottery as an instrument to study the impact of military service on earnings.
- **Propensity Score Matching (PSM)**:
  - Aims to match treatment and control units based on a calculated propensity score to account for observed covariates.
  - Example: Evaluating the impact of online ad campaigns by matching users based on their likelihood of seeing ads.
- **Difference in Differences (DID)**:
  - Measures the effect of a treatment by comparing the changes in outcomes over time between a treatment group and a control group.
  - Example: Assessing the impact of TV advertising on user engagement by comparing geographical areas with and without ads.

12 Client Side Experiments

Problem Statement

The chapter addresses the complexities and implications of running experiments on client-side applications (thick clients) such as mobile apps and desktop software, compared to server-side (thin clients) like web browsers. The focus is on understanding how the release process, data communication, and user behavior differences between these platforms affect the design and analysis of experiments.

Motivating Example

Consider a scenario where a social media app decides to test a new feature that changes how notifications are displayed. The feature's impact on user engagement needs to be assessed through an experiment. However, unlike web-based services where updates are immediate, the mobile app requires navigating app store approvals, varied user update behaviors, and different device capabilities. This example illustrates the challenges in deploying and measuring the effectiveness of new features in thick client environments.

Solutions and Implications

- **Release Process Challenges:**
  - *Controlled Rollouts:* Use of staged rollout features in app stores to manage deployment and mitigate risks.
  - *Version Disparity:* Handling multiple live app versions due to asynchronous user updates.
- **Data Communication:**
  - *Connectivity Issues:* Design experiments considering potential delays in data sync due to inconsistent internet access.
  - *Resource Constraints:* Account for device limitations like battery life, CPU, and data usage which can affect app performance and user engagement.
- **Experiment Design Considerations:**
  - *Parameterization:* Embedding experiments within app configurations to allow flexibility in feature testing without frequent updates.
  - *Delayed Experiment Start:* Anticipate delays in experiment effectiveness due to the time it takes users to update the app.
- **Operational Implications:**

  - *Failsafe Mechanisms:* Implement default settings for experiments to handle offline scenarios or server communication failures.
  - *Health Metrics:* Monitor device and app-level metrics like battery usage and crash rates to understand indirect impacts of experiments.

- **Analytical Challenges:**
  - *Biased Data:* Adjust for biases due to different user update behaviors and device capabilities.
  - *Multi-platform Interactions:* Consider user behavior across different devices and platforms to accurately measure experiment impacts.

13 Instrumentation

Introduction

Instrumentation is essential for understanding user interactions and system performance in any technological setup, particularly in web and application environments. It involves tracking and logging user activities like clicks and hovers, as well as system responses such as latencies and error rates. This chapter emphasizes the importance of implementing effective instrumentation strategies to gather meaningful data, which is crucial for running successful experiments and enhancing user experience.

Client-Side vs. Server-Side Instrumentation

- **Client-Side Instrumentation:**
  - Focuses on user experience—actions (clicks, hovers), performance (page load times), and errors (JavaScript errors).
  - Drawbacks include potential negative impacts on user experience due to increased load times and battery usage.
  - Data accuracy issues due to lossy JavaScript instrumentation, with specific scenarios where data loss occurs during page transitions.

- **Server-Side Instrumentation:**
  - Concentrates on system performance—response times, request handling, and error logging.
  - Provides more reliable and granular data which is less susceptible to the variances seen in client-side data.
  - Useful for internal diagnostics like debugging and tuning system algorithms.

Processing Logs from Multiple Sources

- Ensuring logs from various sources (client types, servers) are combinable and useful for downstream processing.
- Importance of a common identifier or join key across logs to correlate events and user actions for comprehensive analysis.
- Shared formats and common fields in logs facilitate easier and more effective data analysis and segmentation.

Culture of Instrumentation

- Emphasizes the critical nature of robust instrumentation, likening inadequate instrumentation to flying a plane with faulty instruments.
- Advocates for a cultural shift where instrumentation is integral to development, not an afterthought.
- Suggests practices such as including instrumentation in the initial spec, testing during development, and monitoring log quality to ensure data integrity and usefulness.

14 Choosing A Randomization Unit

Problem Statement

The chapter addresses the critical issue of selecting an appropriate randomization unit in experimental design, which impacts both user experience and the effectiveness of metric evaluation in experiments.

Motivating Example

The historical example from RAND (1955) illustrates the challenges in generating unbiased random digits and the necessity of refining randomization processes. This underscores the importance of choosing the right randomization unit to ensure the integrity and reliability of experimental outcomes.

Solutions and Considerations

- **Granularity Considerations:**
  - *Page-level:* Each page view is a unit. High granularity but may lead to inconsistent user experiences if the treatment varies from page to page.
  - *Session-level:* All pages viewed in a single session are considered one unit. Balances between granularity and user experience consistency.
  - *User-level:* All user interactions are treated as a single unit. Provides consistency but may reduce the number of randomization units available.
- **Impact on Metrics:**
  - Finer granularity increases the number of units, enhancing statistical power but potentially affecting cross-page or cross-session features.
  - Coarser granularity ensures better user experience consistency and aligns better with user-level metrics.
- **Stable Unit Treatment Value Assumption (SUTVA):**
  - Ensuring that the treatment of one unit does not affect the treatment of another is crucial, especially in finer granularities where user perception might lead to behavior changes.
- **Randomization and Analysis Alignment:**
  - The randomization unit should ideally be the same as or coarser than the analysis unit to simplify the analysis and ensure accurate variance estimation.
- **User-level Randomization Options:**
  - *Signed-in user ID:* Offers cross-device consistency and longitudinal stability.
  - *Pseudonymous ID (e.g., cookies):* Less stable but useful for non-signed-in user scenarios.
  - *Device ID:* Stable but limited to specific devices, lacking cross-device tracking capabilities.

15 Ramping Experiment Exposure

Problem Statement

The chapter addresses the challenge of balancing speed, quality, and risk in the process of ramping up experiment exposure for new feature launches in controlled online environments.

Motivating Example

A notable example discussed is the initial launch of Healthcare.gov, which faced significant issues due to full exposure on day one without incremental ramping. This led to system overload and failure, highlighting the critical need for a controlled ramping process to mitigate risks and ensure system readiness.

Solutions and Techniques

- **Pre-MPR (Minimum Power Ramp)**:
  - Utilize "rings" of testing populations to gradually increase exposure and mitigate risks.
  - Implement automated traffic dial-up to desired allocations, enhancing control over exposure levels.
  - Employ real-time monitoring of key metrics to quickly identify and address potential issues.
- **MPR (Maximum Power Ramp)**:
  - Maintain experiments at MPR typically for a week to account for time-dependent factors and ensure reliable measurements.
  - Adjust duration based on the presence of novelty or primacy effects which might skew the results.
- **Post-MPR**:
  - Focus on operational concerns and infrastructure readiness, with short, closely monitored ramps.
- **Long-Term Holdout or Replication**:
  - Consider long-term effects and sustainability of the treatment.
  - Use holdouts to measure cumulative impacts and validate surprising results through replication.

16 Scaling Experiment Analyses

Problem Statement

The chapter addresses the challenge of scaling data analysis pipelines within experimentation platforms to ensure robust, consistent, and scientifically sound methodologies that are also trustworthy and efficient.

Motivating Example

A company aiming to transition to advanced stages of experimentation maturity needs to integrate robust data processing and computation frameworks. For instance, a scenario where user interaction data from various logs are sorted, cleaned, and enriched to provide reliable inputs for experiment analysis illustrates the necessity of a systematic approach to handle large-scale data efficiently.

Data Processing Solutions

- **Sorting and Grouping:**
  - Sort by user ID and timestamp to facilitate session creation and event grouping.
  - Virtual joins may suffice, avoiding the need for materializing joins unless for broader uses like debugging or hypothesis generation.
- **Cleaning the Data:**
  - Apply heuristics to identify and remove non-human interactions such as bots.
  - Address data quality issues like duplicate events or incorrect timestamps.
- **Enriching the Data:**
  - Enhance data with additional metrics such as browser type or session duration, crucial for detailed experiment analysis.
  - Annotate data to flag its relevance to specific experiments, optimizing computation resources.

Data Computation Approaches

- **Materialized Per-User Statistics:**
  - Useful for both business reporting and specific experiments.
  - Facilitates efficient use of compute resources by reusing user-level aggregated data.
- **Integrated Experiment-Specific Computation:**
  - Metrics and segments are computed on-the-fly, tailored to specific experiments.
  - Ensures flexibility and resource efficiency but requires rigorous consistency checks across pipelines.

Results Summary and Visualization

- Highlight critical metrics and statistical significance using visual aids like color-coding.
- Provide tools for segment drill-downs to explore detailed impacts and improve decision-making.
- Ensure the visualization tools cater to a diverse audience, enhancing accessibility and understanding across different organizational roles.

17 Statistics Behind Online Controlled Experiments

Problem Statement

This chapter delves into the statistical intricacies essential for designing and analyzing online controlled experiments, focusing on hypothesis testing, statistical power, and the assumptions underlying these tests.

Motivating Example

Consider an online platform testing a new feature aimed at increasing user engagement measured by queries-per-user. A two-sample t-test is employed to determine if the observed increase in engagement for the treatment group (users exposed to the new feature) compared to the control group (users not exposed) is statistically significant or merely due to random variation.

Detailed Solutions

- **Two-Sample t-Test**
  - Hypotheses: H_0: mean(Y_t) = mean(Y_c), H_A: mean(Y_t) $\neq$ $mean($Y_c$).t-$ $Statistic : $T = \frac{\Delta}{\sqrt{\text{var}(\Delta)}}$, $where$\Delta = Y_t - Y_c$.
- Significance is determined by the p-value; a p-value ¡ 0.05 typically indicates a statistically significant difference.
- **p-Value and Confidence Interval**
  - p-Value: Probability of observing a test statistic as extreme as T, or more, under the null hypothesis.
  - Confidence Interval: If the 95% confidence interval for $\Delta$$doesnotincludezero, theresultisdeemedsignificantatthe0
- **Normality Assumption**
  - Central Limit Theorem justifies the normality of $$ for large sample sizes, despite non-normal distribution of individual Y values.
  - Minimum sample size for normal approximation can be estimated using skewness of the distribution.
- **Type I/II Errors and Power**

  - Type I error (false positive): Concluding a significant difference when there is none; controlled at a rate of 0.05.
  - Type II error (false negative): Failing to detect a true difference; inversely related to power.
  - Power: Probability of correctly rejecting H_0 when H_A is true, typically set at 80%.

- **Bias and Multiple Testing**
  - Bias: Systematic difference between the estimated and true values, can be due to various experimental flaws.
  - Multiple Testing: Adjusting significance thresholds using methods like Bonferroni correction to control the family-wise error rate.

- **Meta-Analysis**
  - Combining p-values from multiple independent tests to increase power and reduce false positives using Fisher's method.

18 Variance Estimation And Improved Sensitivity

Problem Statement

This chapter addresses the critical issue of variance estimation in statistical analysis, which is fundamental for computing p-values and confidence intervals. Incorrect estimation can lead to false negatives or positives, affecting the reliability of hypothesis tests.

Motivating Example

Consider an online platform testing a new feature intended to increase user engagement. The platform measures the average session duration per user, a metric susceptible to high variance due to user behavior diversity. Accurate variance estimation is crucial to determine if observed changes are due to the new feature or random fluctuations.

Common Pitfalls and Solutions

- **Delta vs. Delta %**
  - Incorrect variance estimation for relative differences can mislead decision-makers about the impact of test results.
  - Correct approach: Estimate variance of the ratio, not just the difference, to account for the variability in the denominator.

- **Ratio Metrics**
  - Challenge arises when the analysis unit differs from the experiment unit, leading to potential correlation issues.
  - Solution: Use the delta method for ratio metrics to estimate variance correctly, ensuring the analysis respects the underlying distribution assumptions.

- **Outliers**
  - Outliers can disproportionately affect variance estimates, leading to misleading test results.
  - Practical approach: Cap extreme values or use robust statistical methods to mitigate the impact of outliers.

Improving Sensitivity

- **Metric Transformation**
  - Use transformations like log or binary indicators to reduce variance and improve the sensitivity of the tests.
- **Stratification and Control Variates**
  - Implement stratification by dividing data into homogeneous subgroups or use covariates to adjust for known sources of variability.
- **Granular Randomization**
  - Increase the granularity of randomization (e.g., per page or query) to enhance the precision of variance estimates.
- **Paired Designs**
  - Use paired experimental designs to directly compare treatment and control conditions within the same subjects, reducing variability.
- **Pooling Control Groups**
  - Combine control groups from multiple concurrent experiments to increase the overall sample size and statistical power.

Variance of Other Statistics

- For non-average statistics like quantiles, consider methods like bootstrap or density estimation to handle the unique challenges posed by these metrics.

19 The Aa Test

Problem Statement

The chapter delves into the importance and methodology of A/A tests in the context of controlled experiments, particularly highlighting their role in verifying the reliability of an experimentation platform. A/A tests are designed to ensure that any statistical significance in experiment outcomes is due to the experimental treatment and not due to any inherent biases or errors in the experimental setup.

Motivating Example

Consider a scenario where a company wants to evaluate the performance of a new website design against the old one. Before running a direct A/B test, they conduct an A/A test where both groups are exposed to the old website. This preliminary step helps to confirm that the system behaves as expected under controlled conditions, ensuring that any subsequent differences observed in the A/B test can be attributed to the new design and not to other confounding factors.

Detailed Solutions

- **Continuous Testing:** Regularly run A/A tests alongside A/B tests to detect and correct any discrepancies in data handling or experiment execution.
  - Helps in identifying platform level biases or carry-over effects from previous experiments.
- **Statistical Analysis:**
  - Use t-tests and analyze the distribution of p-values. A uniform distribution of p-values indicates that the system is free from biases.
  - Apply corrections like the delta method if the p-value distribution is skewed, indicating potential issues with variance estimates.
- **Data Integrity Checks:**

  - Compare key metrics like user counts and revenue with the system of record to ensure data consistency.
  - Check for user leakage or discrepancies in data collection methods.
- **Variance Estimation:** Use data from A/A tests to estimate the variance of metrics, which aids in calculating the statistical power and determining the required duration for A/B tests.
- **Simulation Techniques:**
  - Simulate multiple A/A tests using historical data to predict and rectify potential issues in real-time data handling.
  - This approach helps in fine-tuning the experimentation platform before deploying actual A/B tests.

20 Triggering For Improved Sensitivity

Problem Statement

The chapter addresses the problem of improving statistical power in experiments by filtering out noise from users who could not have been impacted by the experiment. This is achieved through a method called "triggering," which involves analyzing only those users who could potentially show a difference due to the experiment.

Motivating Example

Consider an e-commerce site testing a new checkout process. Only users who initiate the checkout process are relevant for the analysis because the treatment effect for users not initiating checkout is zero. Analyzing all users would dilute the statistical power by including noise from those unaffected by the change.

Solutions and Methods

- **Intentional Partial Exposure:** Analyze only users from a specific segment affected by the change, e.g., only US users if the experiment is run in the US.
  - Include mixed users if they could have been exposed to the change.
- **Conditional Exposure:** Trigger users into the experiment only if they use the feature or reach the part of the site being tested.
  - Example: Only trigger users who start the checkout process if the checkout process is changed.
- **Coverage Increase:** Trigger only those users who meet specific new criteria introduced by the treatment.
  - Example: In a free shipping offer experiment, only trigger users whose cart value falls within the newly set range.
- **Coverage Change:** Adjust the triggering based on changes in coverage of the treatment.
  - Example: If free shipping is offered to a different cart value range, trigger users based on the new and old criteria.
- **Counterfactual Triggering for Machine Learning Models:** Trigger users based on whether the output of a new model differs from the control model.
  - This requires running both models (control and treatment) concurrently and comparing outputs in real-time.
- **Optimal and Conservative Triggering:**
  - Optimal: Trigger only users showing a difference due to the treatment.
  - Conservative: Include more users than optimal to simplify the process, though this may reduce statistical power.

21 Sample Ratio Mismatch And Other Trust Related Issues

Problem Statement

The chapter focuses on the problem of Sample Ratio Mismatch (SRM) and its impact on the trustworthiness and validity of experimental results in statistical and machine learning contexts.

Motivating Example

In a controlled experiment with a 50% assignment to both Control and Treatment groups, an unexpected SRM was observed: Control had 821,588 users while Treatment had 815,482 users, yielding a ratio of 0.993 against the designed 1.0. The p-value of this mismatch was extremely low (1.8E-6), indicating a highly unlikely event under the assumption of correct experimental execution. This suggests a potential bug in the experiment's implementation, leading to distrust in all other metrics derived from the experiment.

Solutions

- **Check Randomization and Assignment:**
  - Ensure that the randomization process upstream of the experiment is not biased.
  - Verify that variant assignments are correct and that users are properly randomized at the start of the data pipeline.
- **Analyze Data Pipeline Integrity:**
  - Investigate each stage of the data processing pipeline for potential sources of SRM, such as bot filtering or caching issues.
  - Consider the impact of concurrent experiments and ensure isolation groups are correctly managed.
- **Segment Analysis:**
  - Examine the sample ratio daily and across different user segments (e.g., browser type, new vs. returning users) to identify specific points or segments where mismatches occur.
- **Debugging and Corrective Actions:**
  - If an SRM is detected, prioritize debugging over analyzing other metrics. Use insights from the SRM analysis to identify and correct the underlying issues.
  - Re-run the experiment if necessary, after addressing the identified issues to ensure the integrity and trustworthiness of the results.

22 Leakage And Interference Between Variants

Problem Statement

This chapter addresses the problem of leakage and interference between variants in controlled experiments, particularly when the Stable Unit Treatment Value Assumption (SUTVA) fails. SUTVA assumes that the behavior of each unit in an experiment is unaffected by the variant assignment to other units, which is not always the case.

Motivating Example

Consider a social network where user behavior is influenced by their connections. If a new feature is tested, such as a video chat on Facebook, and if it becomes popular among a user's friends, that user is more likely to adopt it too. This creates a scenario where the behavior of users in the control group is influenced by the behavior of users in the treatment group, leading to interference. This spillover effect can skew the results of the experiment, making it difficult to isolate the impact of the new feature.

Solutions

- **Rule-of-Thumb: Ecosystem Value of an Action**
  - Identify actions with potential spillover effects and measure their ecosystem impact.
  - Use historical data to establish a baseline for expected impacts using the Instrumental Variable approach.
- **Isolation Techniques**
  - **Splitting Shared Resources:** Allocate resources like ad budgets or training data specifically to treatment and control to prevent interference.
  - **Geo-based Randomization:** Use geographical separation to minimize interference between units.
  - **Time-based Randomization:** Assign treatment and control at different times to prevent simultaneous interference.
  - **Network-cluster Randomization:** Randomize clusters of closely connected units to minimize edge effects between clusters.
  - **Network Ego-centric Randomization:** Focus on isolating 'ego' nodes and their direct connections to control spillover within a network.
- **Edge-Level Analysis**
  - Analyze interactions between users based on their treatment assignment to understand network effects and biases.
- **Detecting and Monitoring Interference**
  - Implement robust monitoring systems to detect and alert for potential interference during experiments.
  - Use ramp-up phases to identify and mitigate severe interference effects before full deployment.

23 Measuring Long Term Treatment Effects

Problem Statement

The chapter addresses the challenge of measuring long-term treatment effects in environments where products and services evolve rapidly and are often measured only in the short term. This is crucial as long-term effects can significantly differ from short-term outcomes, affecting strategic decisions and product development.

Motivating Example

Consider the scenario of an online platform that introduces a new feature aimed at increasing user engagement. Initially, the feature may show positive effects by increasing user interactions. However, over time, users might experience fatigue, reducing engagement levels below the initial baseline. This example illustrates the importance of understanding how the initial gains can transform into long-term outcomes.

Solutions and Methods

- **Long-Running Experiments:**
  - Measure the treatment effect at both the start and end of the experiment.
  - Challenges include treatment effect dilution, multiple device usage, and cookie churn impacting the accuracy of long-term effect measurements.
- **Cohort Analysis:**
  - Analyze a stable cohort over time to control for dilution and survivorship bias.

  - Considerations include ensuring cohort stability and representativeness to avoid external validity issues.

- **Post-Period Analysis:**
  - After ending the treatment, measure the ongoing effects to understand user or system-learned behaviors.
  - This method helps isolate the learned effects from other variables.

- **Time-Staggered Treatments:**
  - Use staggered treatment start times to measure when treatment effects stabilize.
  - This method assumes that the difference between staggered treatments decreases over time.

- **Holdback and Reverse Experiment:**
  - Maintain a control group even after widespread deployment to measure long-term effects against a baseline.
  - Reverse experiments reintroduce control conditions to previously treated users to assess changes upon withdrawal.

24 References

Problem Statement

The chapter compiles a vast array of references spanning various aspects of statistics, machine learning, and data-driven decision-making. It aims to provide a comprehensive resource for understanding the evolution and application of statistical methods and machine learning in experimental design and analysis.

Motivating Example

The references include seminal works such as Fisher's 1925 introduction of statistical methods for research, which revolutionized how scientific experiments are conducted and interpreted. This foundational work underpins modern experimental design, influencing methodologies in fields ranging from agriculture to medicine, and now prominently in tech industries where A/B testing and controlled experiments shape product development and user experience strategies.

Solutions and Methods

- **Statistical Inference and Hypothesis Testing:**
  - Works by Casella and Berger, and Efron and Tibshirani provide insights into statistical inference, emphasizing the importance of hypothesis testing and the bootstrap method.
  - Benjamini and Hochberg's procedures for controlling the false discovery rate address multiple comparisons problems in hypothesis testing.

- **Experimental Design:**
  - Box, Hunter, and Hunter's texts guide the design, innovation, and discovery phases of experiments, crucial for effective implementation in industrial and scientific research.
  - Kohavi et al.'s papers discuss the nuances of online controlled experiments, particularly in web environments, highlighting challenges like selection bias and data snooping.

- **Machine Learning and Data Mining:**
  - Athey and Imbens focus on machine learning approaches for causal inference, essential for understanding treatment effects in experiments.
  - Chapelle et al. explore large-scale validation in search engine algorithms, pivotal for refining user interaction models based on implicit feedback.

- **Ethics and Data Privacy:**

  - Dwork and Roth's exploration of differential privacy frameworks is critical for conducting data analysis while safeguarding user privacy.
  - Articles on the ethics of online research highlight the delicate balance between innovation and ethical responsibility in experimental design.

- **Advanced Statistical Techniques:**
  - Articles on Bayesian methods and modern approaches to regression and classification provide advanced tools for data analysis, enhancing the precision of experimental outcomes.
  - Techniques for handling large datasets, as discussed by McFarland and colleagues, are crucial for the scalability of data-driven methodologies in business and technology sectors.