Lecture 10

Ksenia Kasianova

Plan

A/B testing

Uber A/B/N model

Classic A/B testing

Analysis Steps

Key components and statistical methodologies

Delta method

Metrics management

Sequential testing

Continuous experi-

Lecture 10: A/B testing: summary

Lecturer: Ksenia Kasianova
xeniakasianova@gmail.com

January 29, 2024

Plan

1) A/B/N testing

2) Sequential testing

3) Continuous testing

A/B testing

Lecture 10

Ksenia Kasianova

Plan

A/B testing

Uber A/B/N model

Classic A/B testing

Analysis Steps

Key components and statistical methodologies

Delta method

Metrics management

Sequential testing

Continuous experi-

Where does a company conduct A/B tests?

Using A/B testing:

— Experimentally test mathematical and statistical models.

— Optimize communication with users so as not to spam them.

— Offline tests of recommendation models.

For example, send information about new content that may interest them to users.

— Conducts online tests of recommendation models.

A/B testing

Lecture 10

Ksenia Kasianova

Plan

A/B testing

Uber A/B/N model

Classic A/B testing

Analysis Steps

Key components and statistical methodologies

Delta method

Metrics management

Sequential testing

Continuous experi-

Each company has its own specifics, on which preferred methodology depends:

— Small data sets

When there is no history, experiments are carried out from scratch.

— Conduct experiments on groups of real users.

Sometimes the groups are small, which may introduces bias into the estimation results.

— Time of response

E.g. Communication with the user may take place once a week.

— Conduct experiments that are timed to coincide with the release of specific content.

No counterfactual (no copy of Russia where content was not released)

— Seasonality

There are metrics that are highly shifted, so the results of the experiment cannot always be reproduced.

E.g. on January holidays, users watch more content than on weekdays.

Uber A/B/N model

Lecture 10

Ksenia Kasianova

Plan

A/B testing

Uber A/B/N model

Classic A/B testing

Analysis Steps

Key components and statistical methodologies

Delta method

Metrics management

Sequential testing

Continuous experi-

Goals: launch, debug, measure, and monitor the effects of

– new ideas

– product features

– marketing campaigns

– promotions,

– and even machine learning models.

$=>$ **Experiments** across driver, rider, Uber Eats, and Uber Freight apps using

– A/B/N

– causal inference

– multi-armed bandit (MAB)-based continuous experiments.

Uber A/B/N model

Lecture 10

Ksenia Kasianova

Plan

A/B testing

Uber A/B/N model

Classic A/B testing

Analysis Steps

Key components and statistical methodologies

Delta method

Metrics management

Sequential testing

Continuous experi-

Main types of statistical methodologies:

– fixed horizon A/B/N tests (t-test, chi-squared, and rank-sum tests),

– sequential probability ratio tests (SPRT),

– causal inference tests (synthetic control and diff-in-diff tests),

– and continuous A/B/N tests using bandit algorithms (Thompson sampling, upper confidence bounds, and Bayesian optimization with contextual multi-armed-bandit tests).

To estimate standard errors:

– block bootstrap

– delta methods

To measure bias correction when calculating the probability of type I and II errors:

– regression-based methods

Uber A/B/N model

Lecture 10

Ksenia Kasianova

Plan

A/B testing

Uber A/B/N model

Classic A/B testing

Analysis Steps

Key components and statistical methodologies

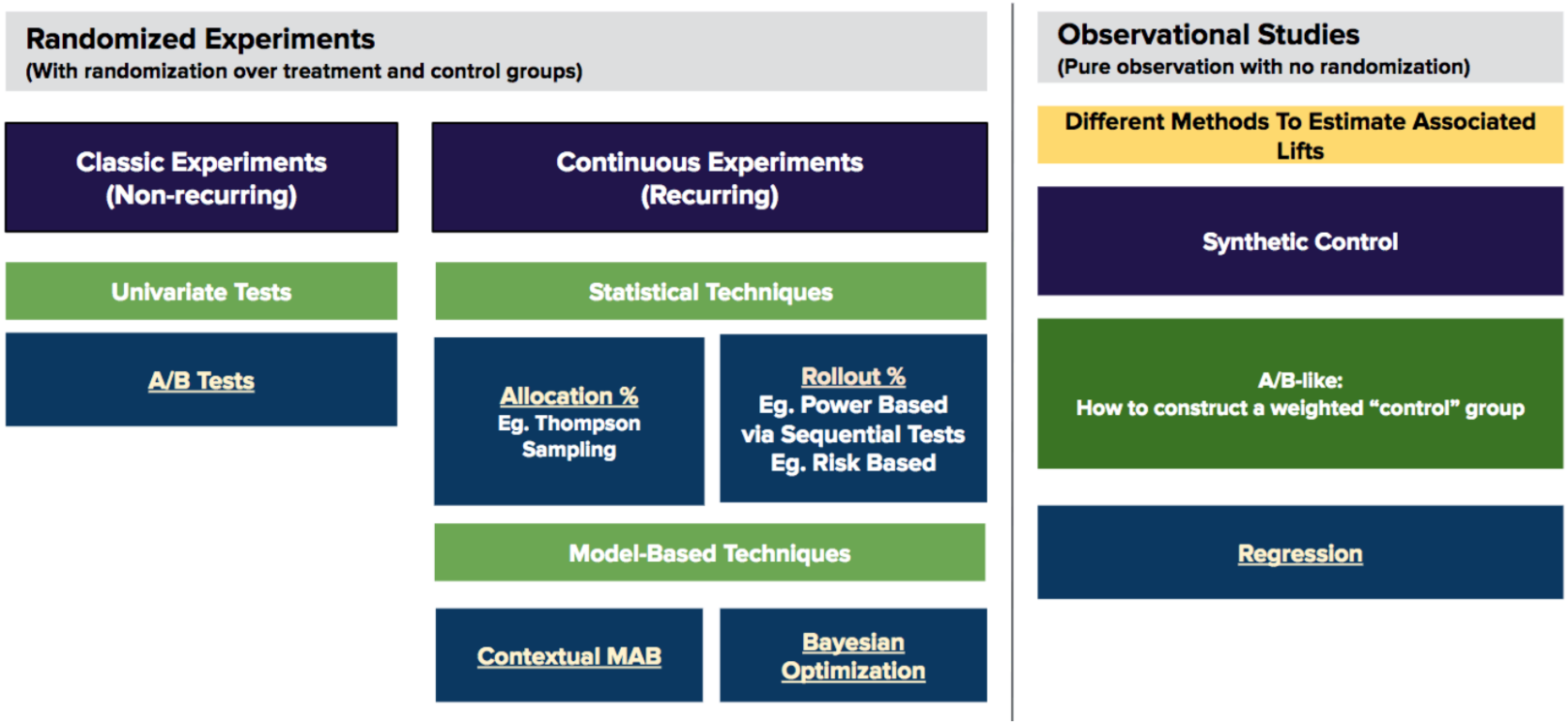
Delta method

Metrics management

Sequential testing

Continuous experi-

Overview of data generation, modeling and interpretation in statistical perspectives



Classic A/B testing

Lecture 10

Ksenia Kasianova

Plan

A/B testing

Uber A/B/N model

Classic A/B testing

Analysis Steps

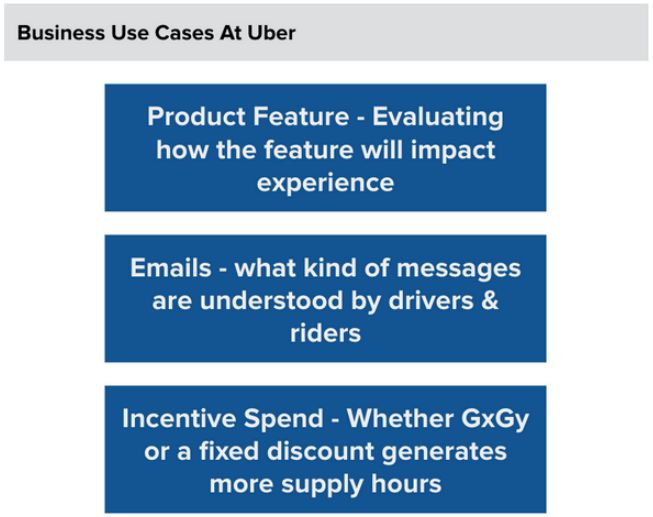
Key components and statistical methodologies

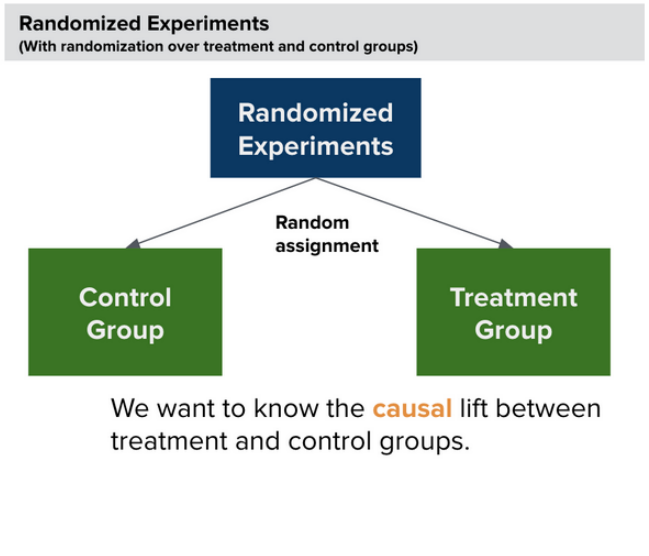
Delta method

Metrics management

Sequential testing

Continuous experi-





Common use: feature release experiments.

Analysis Steps

Lecture 10

Ksenia Kasianova

Plan

A/B testing

Uber A/B/N model

Classic A/B testing

Analysis Steps

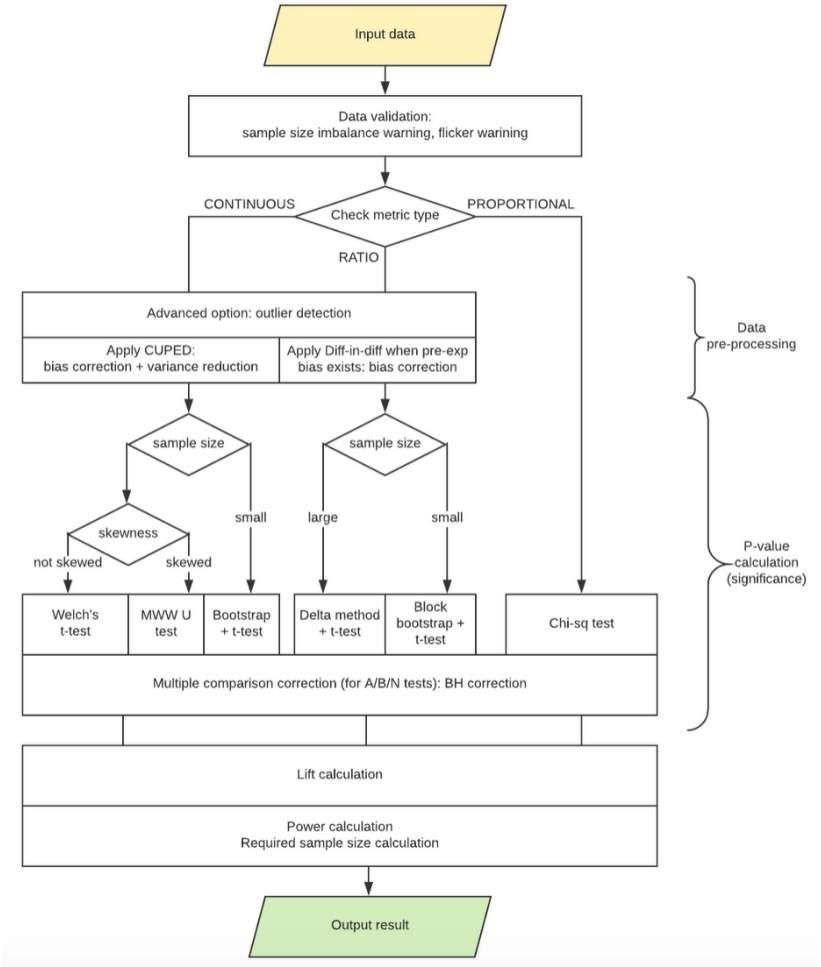
Key components and statistical methodologies

Delta method

Metrics management

Sequential testing

Continuous experi-



Uber's Statistical engine

1) Pick a decision metric (e.g., rider gross bookings).

– This choice relates directly to the hypothesis being tested.

2) Engine reuses pre-defined metrics and automatically handles data gathering and data validation.

3) Select appropriate statistical model

4) Output = easy-to-read report

Key components and statistical methodologies

1) Data validation

Major issues for experimenters to watch for and to keep a *healthy skepticism* in their A/B experiments:

– Sample size imbalance

i.e. sample size ratio in the control and treatment groups is significantly different from what was expected.

Warning: double check their randomization mechanisms.

– Flickers

i.e. users that have switched between control and treatment groups.

Most of our use cases are randomized experiments and most of the time summarized data is sufficient for performing fixed horizon A/B tests.

Key components and statistical methodologies

2) At the user level, there are three distinct metrics:

– Continuous metrics contain one numeric value column, e.g., gross bookings per user.

– Proportion metrics contain one binary indicator value column, e.g., to test the proportion of users who complete any trips after sign-up.

– Ratio metrics contain two numeric value columns, the numerator values and the denominator values,

e.g., the trip completion ratio, *number of completed trips* vs *number of total trip requests*.

Key components and statistical methodologies

Lecture 10

Ksenia Kasianova

Plan

A/B testing

Uber A/B/N model

Classic A/B testing

Analysis Steps

Key components and statistical methodologies

Delta method

Metrics management

Sequential testing

Continuous experi-

3) Three variants of data preprocessing:

Goal: to improve the robustness and effectiveness of our A/B analyses

– Outlier detection

Goal: to remove irregularities in data and improves the robustness of analytic results

Method: a clustering-based algorithm for outlier detection and removal.

– Variance reduction

Goal: increase the statistical power of hypothesis testing.

Especially important for a) small user base experiments, or (b) when we need to end the experiment prematurely without sacrificing scientific rigor.

Method: CUPED Method leverages extra information we have and reduces the variance in decision metrics.

– Pre-experiment bias

Goal: produce reliable treatment effects estimation by constructing robust counterfactual when mere randomization isn't enough.

Big challenge because of the diversity of users.

Method: Difference in differences (diff-in-diff) is a well-accepted method to correct pre-experiment bias between groups

Key components and statistical methodologies

Lecture 10

Ksenia Kasianova

Plan

A/B testing

Uber A/B/N model

Classic A/B testing

Analysis Steps

Key components and statistical methodologies

Delta method

Metrics management

Sequential testing

Continuous experi-

4) P-value calculation

We compare the p-value to the false positive rate (Type-I error) we desire (0.05) in a common A/B test.

The procedures for p-value calculation:

– Welch's t-test

The default test used for continuous metrics, e.g., completed trips.

– The Mann-Whitney U test,

A nonparametric rank sum test used when there is severe skewness in the data.

– The Chi-squared test, used for proportion metrics, e.g., rider retention rate.

– The Delta method (Deng et al. 2011) and bootstrap methods, used for standard error estimation whenever suitable to generate robust results

Used for (a) ratio metrics or (b) with small sample sizes, e.g., the ratio of trips cancelled by riders.

(!) Multiple comparison correction (the Benjamini-Hochberg procedure) is used when there are two or more treatment groups (e.g., in an A/B/C test or an A/B/N test).

Goal: to control the overall false discovery rate (FDR)

Key components and statistical methodologies

5) Power calculation

Goal: to provide additional information about the level of confidence users should put into their analysis.

An experiment with low power will suffer from high false negative rates (Type-II error) and high FDRs.

In the power calculations Uber XP conducts, a t-test is always assumed.

6) Required sample size calculation

The opposite of a power calculation and estimates how many users are required by the experiment for it to achieve a high power (0.8).

Delta method

Task: we need to construct a confidence interval for an expression that is nonlinear in terms of parameters.

For example, in a polynomial model $y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \varepsilon_i$ we may be interested in the confidence interval for the vertex of the parabola, i.e. for $\left(-\frac{\beta_2}{2\beta_3}\right)$.

Asymptotic theory allows us to solve this problem using the so-called [1] delta method.

Let us have some consistent and asymptotically normal parameter estimate:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \xi,$$

where $\xi \sim N(0, \text{var}(\xi))$.

What distribution does the function of this estimate $g(\hat{\beta})$ have?

Recall the Taylor series expansion of a function in a neighborhood of the point x_0:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + o(x - x_0), x \to x_0.$$

Delta method

Let's apply this expansion to the function of our parameter estimate:

$$g(\hat{\beta}) = g(\beta) + g'(\beta)(\hat{\beta} - \beta) + o(\hat{\beta} - \beta);$$
$$g(\hat{\beta}) - g(\beta) = g'(\beta)(\hat{\beta} - \beta) + o(\hat{\beta} - \beta);$$
$$\sqrt{n}(g(\hat{\beta}) - g(\beta)) = g'(\beta)\sqrt{n}(\hat{\beta} - \beta) + \sqrt{n} \cdot o(\hat{\beta} - \beta).$$

Because

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \xi,$$

where $\xi \sim N\left(0, \sigma^2\right)$, then

$$\sqrt{n}(g(\hat{\beta}) - g(\beta)) \xrightarrow{d} g'(\beta)\xi.$$

According to the variance property:

$$\text{var}\left(g'(\beta)\xi\right) = \left(g'(\beta)\right)^2 \cdot \text{var}(\xi).$$

Hence:

$$\sqrt{n}(g(\hat{\beta}) - g(\beta)) \xrightarrow{d} N\left(0, g'(\beta)^2 \cdot \text{var}(\xi)\right).$$

Therefore, the random variable $g(\hat{\beta})$ will have an asymptotically normal distribution with expectation $g(\beta)$ and variance $g'(\beta)^2 \cdot \text{var}(\xi)/n$.

Delta method

To understand how the delta method works, consider the paired regression model $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$

Let's construct a confidence interval for the function of the coefficient estimate for the variable $g\left(\hat{\beta}_2\right)$:

$$\sqrt{n}\left(\hat{\beta}_2 - \beta_2\right) \xrightarrow{d} N\left(0, \frac{\text{var}\left(\left(x_i - \mu_x\right)\varepsilon_i\right)}{\left(\text{var}\left(x_i\right)\right)^2}\right).$$

Therefore, in our notations:

$$\text{var}(\xi) = \frac{\text{var}\left(\left(x_i - \mu_x\right)\varepsilon_i\right)}{\left(\text{var}\left(x_i\right)\right)^2}.$$

Thus, the random variable $g\left(\hat{\beta}_2\right)$ will have an asymptotically normal distribution with expectation $g\left(\beta_2\right)$ and variance $g'\left(\beta_2\right)^2 \cdot \text{var}\left(\hat{\beta}_2\right) = g'\left(\beta_2\right)^2 \cdot \frac{\text{var}((x_i - \mu_x)\varepsilon_i)}{(\text{var}(x_i))^2 \cdot n}$.

Delta method

Lecture 10

Ksenia Kasianova

Plan

A/B testing

Uber A/B/N model

Classic A/B testing

Analysis Steps

Key components and statistical methodologies

Delta method

Metrics management

Sequential testing

Continuous experi-

Replace the unknown random variables with their estimates and obtain an estimate of the variance and we get the corresponding standard error:

$$\sqrt{g'\left(\hat{\beta}_2\right)^2 \cdot \text{var}\left(\hat{\beta}_2\right)} = \left|g'\left(\hat{\beta}_2\right)\right| \cdot \text{se}\left(\hat{\beta}_2\right).$$

Thus, the 95% asymptotic confidence interval for the value $g(\beta_2)$ will look like:

$$\left(g\left(\hat{\beta}_2\right) - 1.96 \cdot \left|g'\left(\hat{\beta}_2\right)\right| \cdot \text{se}\left(\hat{\beta}_2\right), \quad g\left(\hat{\beta}_2\right) + 1.96 \cdot \left|g'\left(\hat{\beta}_2\right)\right| \cdot \text{se}\left(\hat{\beta}_2\right)\right).$$

Delta method

Example:

Using the delta method for pairwise regression Estimation of model parameters using OLS allowed us to obtain the following results:

$$\underset{(0,2)}{\hat{y}_i = 2,3 + 4,0x_i}.$$

Construct a 95% asymptotic confidence interval for the value $(\beta_2)^3$.

Solution:

$$\left(g\left(\hat{\beta}_2\right) - 1.96 \cdot g'\left(\hat{\beta}_2\right) \cdot \text{se}\left(\hat{\beta}_2\right), \quad g\left(\hat{\beta}_2\right) + 1.96 \cdot g'\left(\hat{\beta}_2\right) \cdot \text{se}\left(\hat{\beta}_2\right)\right)$$

$$\left(\left(\hat{\beta}_2\right)^3 - 1,96 \cdot 3 \cdot \left(\hat{\beta}_2\right)^2 \cdot \text{se}\left(\hat{\beta}_2\right), \quad \left(\hat{\beta}_2\right)^3 + 1,96 \cdot 3 \cdot \left(\hat{\beta}_2\right)^2 \cdot \text{se}\left(\hat{\beta}_2\right)\right)$$

$$\left(4^3 - 1,96 \cdot 3 \cdot 4^2 \cdot 0,1, \quad 4^3 + 1,96 \cdot 3 \cdot 4^2 \cdot 0,1\right)$$

$$(54,6, \quad 73,4)$$

Metrics management

Lecture 10

Ksenia Kasianova

Plan

A/B testing

Uber A/B/N model

Classic A/B testing

Analysis Steps

Key components and statistical methodologies

Delta method

Metrics management

Sequential testing

Continuous experi-

Q: How to determine the proper metrics to evaluate the performance of an experiment.

A: Collaborative filtering methods used for content recommendation:

– item-based methods (primarily used)

E.g.: if an experimenter switches to the Uber Eats team from the Rider team, it's not necessary for the algorithm to review the previous, Uber Eats-inspired choices of that experimenter when selecting metrics to evaluate.

– user-based methods.

Metrics management

Lecture 10

Ksenia Kasianova

Plan

A/B testing

Uber A/B/N model

Classic A/B testing

Analysis Steps

Key components and statistical methodologies

Delta method

Metrics management

Sequential testing

Continuous experi-

Recommendation engine methodology weighted sum of two scores:

– Popularity score:

The more frequently two metrics are selected together across experiments, the higher the score assigned to their relationship.

Method: Jaccard Index – first known similarity coefficient.

– Absolute score:

Idea: generate a pool of user samples and calculate the Pearson correlation score of the two metrics.

Serendipitous discovery: the experimenter may not have considered adding a metric to the experiment since it is not directly related, but it might be moving with the user-selected metric.

Example:

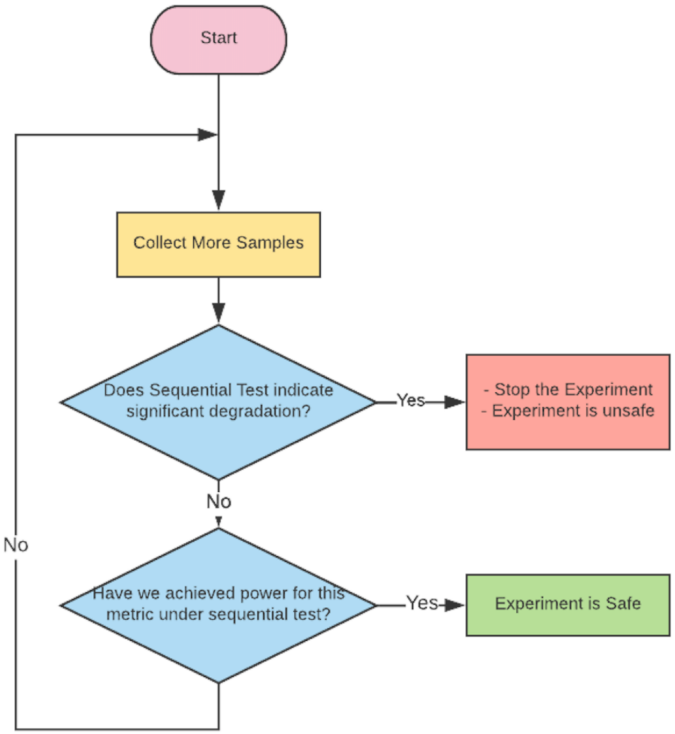
An experimenter wants to measure the treatment effect on driver-partner supply hours

It may not be obvious to the experimenter to also add the number of trips taken by new *riders* as a metric, since this experiment focuses on the *driver* side of the trip equation.

However, both metrics are important for this experiment because of the dynamics of our marketplace.

Sequential testing

Lecture 10
Ksenia Kasianova
Plan
A/B testing
Uber A/B/N model
Classic A/B testing
Analysis Steps
Key components and statistical methodologies
Delta method
Metrics management
Sequential testing
Continuous experi-



Goal: continuously monitor key business metrics

Problem: traditional A/B testing methods (for example, a t-test) inflate Type-I error by repeatedly taking subsamples

Solution: sequential testing

E.g. identifying outages caused by the experiments running on our platform.

Problem:

– cannot wait until a traditional A/B test collects sufficient sample sizes to determine the cause of an outage;

– need to make sure there is no key degradations of business metrics ASAP, during the experimentation period.

Solution: A monitoring system powered by a sequential testing algorithm to adjust the confidence intervals accordingly without inflating Type-I error.

Conduct periodic comparisons about these business metrics, (app crash rates and trip frequency rates), between treatment and control groups for ongoing experiments.

Experiments continue if there are no significant degradations, otherwise they will be given an alert or even paused.

Sequential testing

Lecture 10

Ksenia Kasianova

Plan

A/B testing

Uber A/B/N model

Classic A/B testing

Analysis Steps

Key components and statistical methodologies

Delta method

Metrics management

Sequential testing

Continuous experi-

Mixture Sequential Probability Ratio Test (Johari, 2017)

The test builds on the likelihood ratio test by incorporating an extra specification of mixing distribution H. Suppose we are testing the metric difference with the null hypothesis being θ, then the test statistics could be written as

$$\Lambda_n^{H,\theta_0} = \int_\Phi \Pi_{i=1}^n \frac{f(\theta)(X_i)}{f(\theta_0)(X_i)} h(\theta) d\theta$$

Intuitively, Λ_n^{H,θ_0} represents the evidence against H_0 in favor of a mixture of alternative hypotheses, based on the first n observations.

Sinces we have large sample sizes and the central limit theorem can be applied to most cases, we use normal distribution as our mixing distribution, $H \sim N\left(0, \tau^2\right)$.

This leads to easy computation and a closed form expression for ${\Lambda_n}^{H,\theta_0}$.

Idea: under the null hypothesis, the likelihood ratio at any θ is a martingale, and therefore Λ_n^{H,θ_0} is also a martingale.

Given a desired false positive probability α, it stops and rejects the null hypothesis at the first time $T = T^H(\alpha)$ that $\Lambda_T^{H,\theta_0} \geq \alpha^{-1}$; if no such time exists, it never rejects the null hypothesis.

Using standard martingale techniques (the optional stopping theorem), it can be shown that this sequential test controls Type I error at level α.

Lecture 10

Ksenia Kasianova

Plan

A/B testing

Uber A/B/N model

Classic A/B testing

Analysis Steps

Key components and statistical methodologies

Delta method

Metrics management

Sequential testing

Continuous experi-

Variance estimation with FDR control

To apply sequential testing correctly, we need to estimate variance as accurately as possible.

1) Correlated data

Example: the cumulative difference between our control and treatment groups monitored on a daily basis

Problem: observations from the same users introduce correlations which violate the assumption of the mSPRT test.

e.g. click through rates – the metric from one user across multiple days may be correlated.

Solution to generalize mSPRT test under correlated data:

– delete-a-group jackknife variance estimation

– block bootstrap methods

Sequential testing

2) Multiple hypothesis testing

To evaluate quality of an ongoing experiment many business metrics are monitored at the same time, potentially leading to false alarms.

Theory: Bonferroni or BH correction should be applied

Practice: the potential loss of missing business degradations can be substantial

$\rightarrow$ BH + tuning (MDE, power, tolerance for practical significance, etc.) for metrics with varying levels of importance and sensitivity.

Sequential testing

Lecture 10

Ksenia Kasianova

Plan

A/B testing

Uber A/B/N model

Classic A/B testing

Analysis Steps

Key components and statistical methodologies

Delta method

Metrics management

Sequential testing

Continuous experi-

Suppose we want to monitor a key business metric for a specific experiment using the sequential test methodology

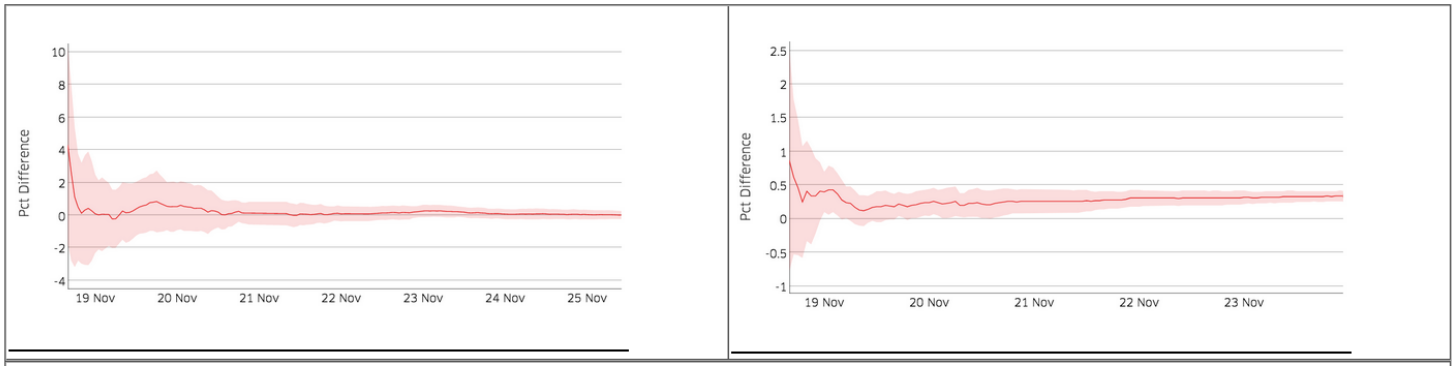


Figure: Plot B – Significant difference between our treatment and control groups. Plot A – no significant difference is identified.

– The red lines Plots A and B – the observed cumulative relative difference between our treatment and control groups.

– The red band – confidence interval for this cumulative relative difference. As time passes, we accumulate more samples and the confidence interval narrows.

In Plot B, the confidence interval consistently deviates from zero starting on a given date, in this example, November 21 $=>$ statistical significance $+$ an extra threshold for practical significance $=>$ metrics degradation is detected after a certain date.

In contrast, Plot A's confidence interval shrinks but always includes 0 $=>$ no regressions for the crash monitored in Plot A.

Continuous experiments

Lecture 10

Ksenia Kasianova

Plan

A/B testing

Uber A/B/N model

Classic A/B testing

Analysis Steps

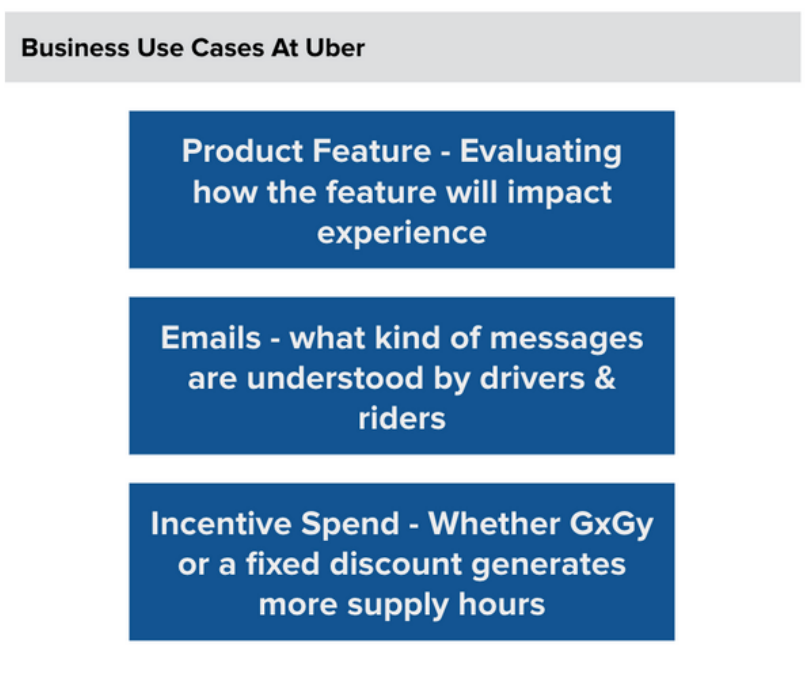
Key components and statistical methodologies

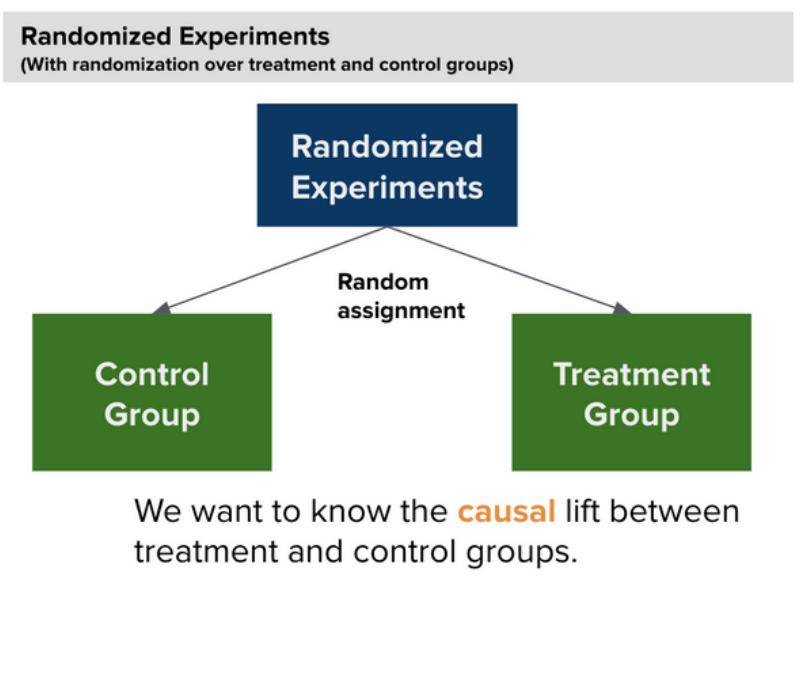
Delta method

Metrics management

Sequential testing

Continuous experiments





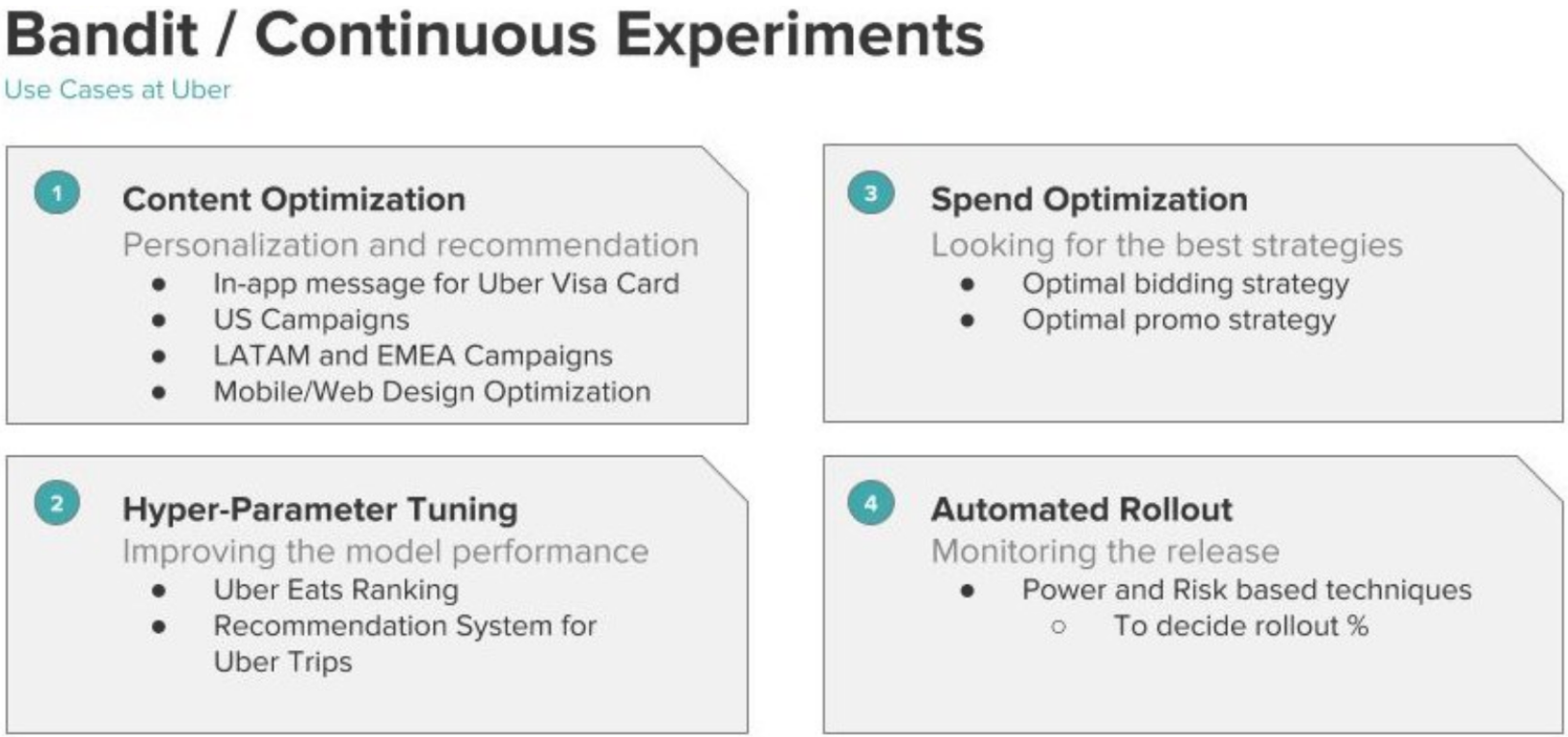
The use of continuous experiments to accelerate innovation and learning:

– bandit,

– and optimization-focused reinforcement learning methods

Goal: learn iteratively and rapidly from the continuous evaluation of related metric performance.

Continuous experiments



Supervised learning – the study of making statistical inferences from previously collected data.

Multi-armed bandits – is more about an interaction between an agent (algorithm) and an environment where one simultaneously collects data and makes inferences in a closed-loop.

Continuous experiments

You have n "arms" or actions, representing distributions. "Pulling" an arm represents requesting a sample from that arm.

At each time $t = 1, 2, 3, \ldots$

- Algorithm chooses an action $I_t \in \{1, \ldots, n\}$

- Observes a reward $X_{I_t,t} \sim P_{I_t}$ where $P_1, \ldots, P_n$ are unknown distributions

That is, playing arm i and time s results in a reward $X_{i,s}$ from the i th distribution.

Goal: to find the most pertinent parameters of these distributions – the means of the distribution

Let $\theta_i^* = \mathbb{E}_{X \sim P_i}[X]$ be the mean of the i th distribution.

Define $\Delta_i = \max_{j=1,\ldots,n} \theta_j^* - \theta_i^*$.

We measure performance of an algorithm in two ways:

1) how much total reward is accumulated, and

2) how many total pulls are required to identify the best mean.

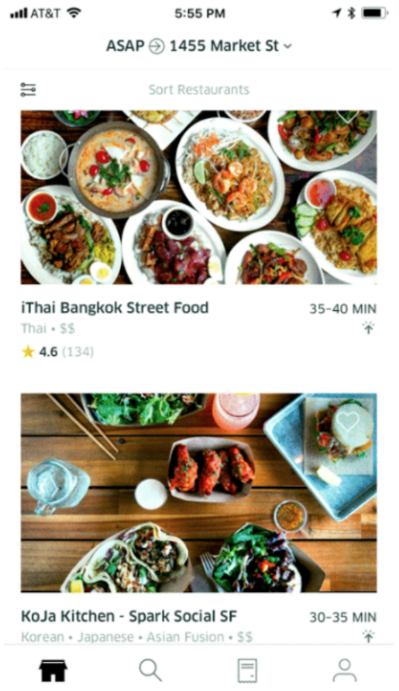
Continuous experiments

Case study: MAB testing to create a linear programming model, called the multiple-objective optimization (MOO), that ranks restaurants on the main feed of the Uber Eats app:

Multiple-objective Optimization (MOO)

Summary of Background Information

- **Topic**
  - Restaurant Recommendation of MainFeed on Uber Eats App
- **Business Problem**
  - Rank the restaurant to drive more revenue, higher conversion and higher retention rates, etc.
- **Modeling Components** [Eats DS Team]
  - **Linear programming**: an optimization problem with multiple objectives
  - **Machine learning**: prediction of business metrics
- **Statistical Methods** [Experimentation Team]
  - **A/B experiments**: testing on the null hypothesis



– many parameter candidates for use with our ranking algorithms.

– ranking results depend on the hyper-parameters we chose for the MOO model.

Continuous experiments

Lecture 10

Ksenia Kasianova

Plan

A/B testing

Uber A/B/N model

Classic A/B testing

Analysis Steps

Key components and statistical methodologies

Delta method

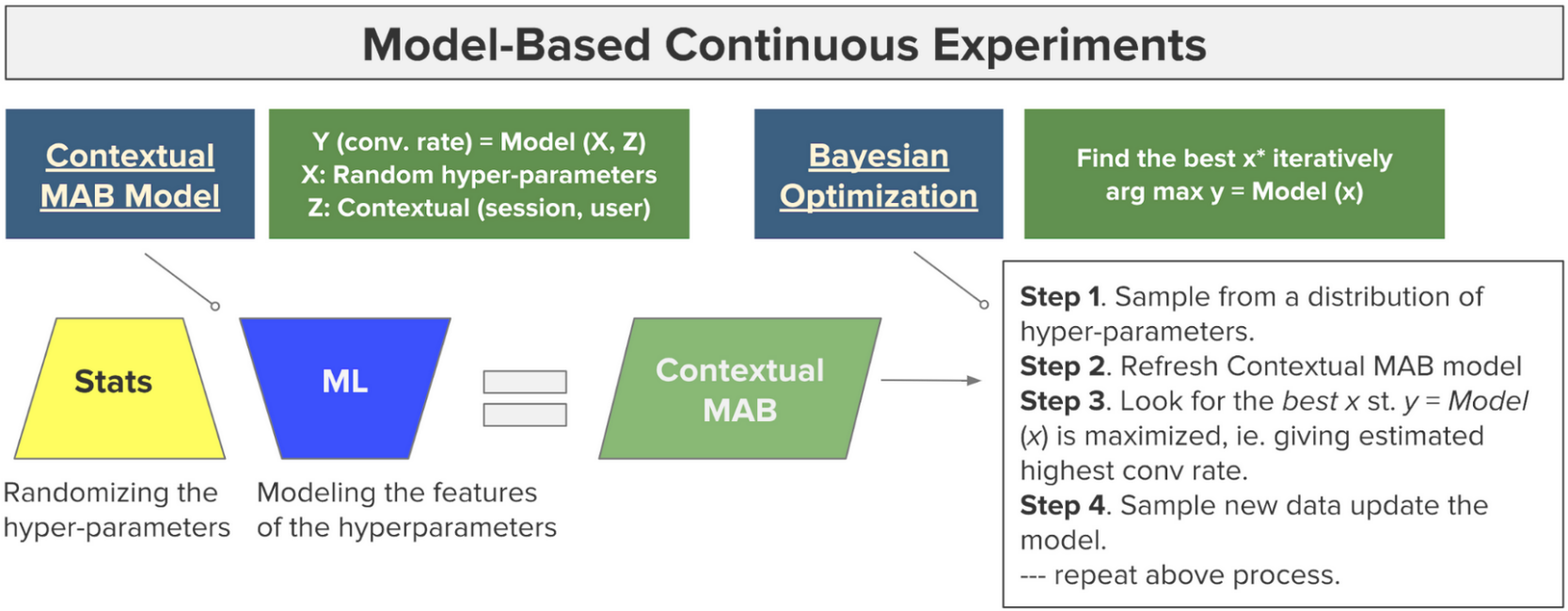
Metrics management

Sequential testing

Continuous experiments

Goal: figure out the best hyper-parameters

The traditional A/B test framework is too time-intensive to handle => multi-armed bandits algorithm.



Summary

1) Algorithm:

– Data validation

– Choice of metrics: null hypothesis, methodology

– Three variants of data preprocessing: Outlier detection, Variance reduction, Pre-experiment bias

– P-value calculation

– Power analysis, sample size calculation

2) Sequential testing vs continuous testing

3) Importance of Communication:

– Listening sessions during the problem discovery and design phase to understand pain points that needed addressing

– Presented new architecture and functionality broadly early on to key customers, explained the customer benefits

– Listening sessions to get feedback on the product – altered our roadmap and reprioritized things to address the pressing needs