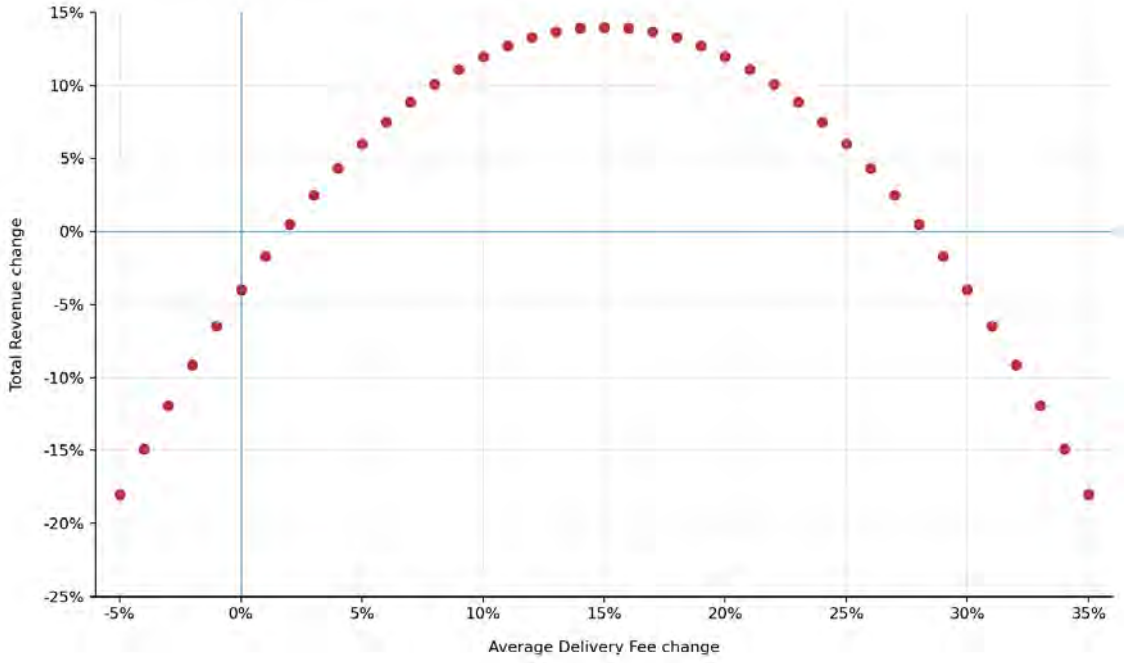
System-Level Experimentation at Scale: Pricing, Supply, A/B and Switchback Testing in Food-Tech

Sergei Nasibian

February 19, 2026

Delivery pricing trade-offs:

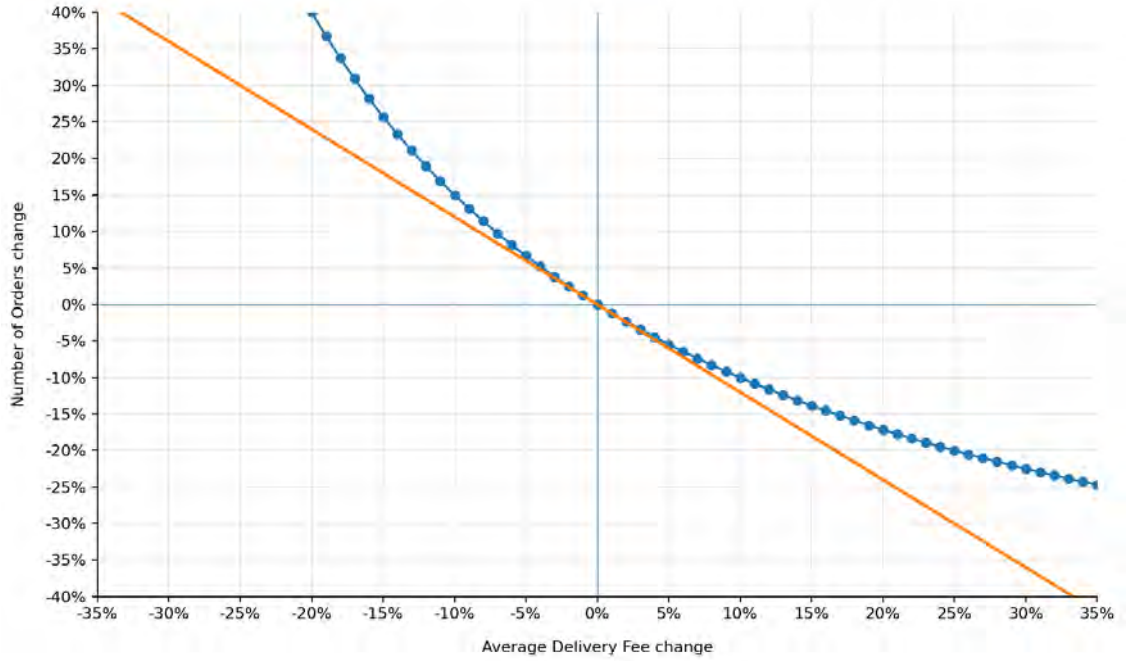
↑ Delivery fee	//	↑ Revenue per order	&	↓ Number of orders
↓ Delivery fee	//	↓ Revenue per order	&	↑ Number of orders



Pricing goals

- **Short-term: total revenue optimisation**
- Long-term: user LTV optimisation

Short-term revenue optimisation: elasticity

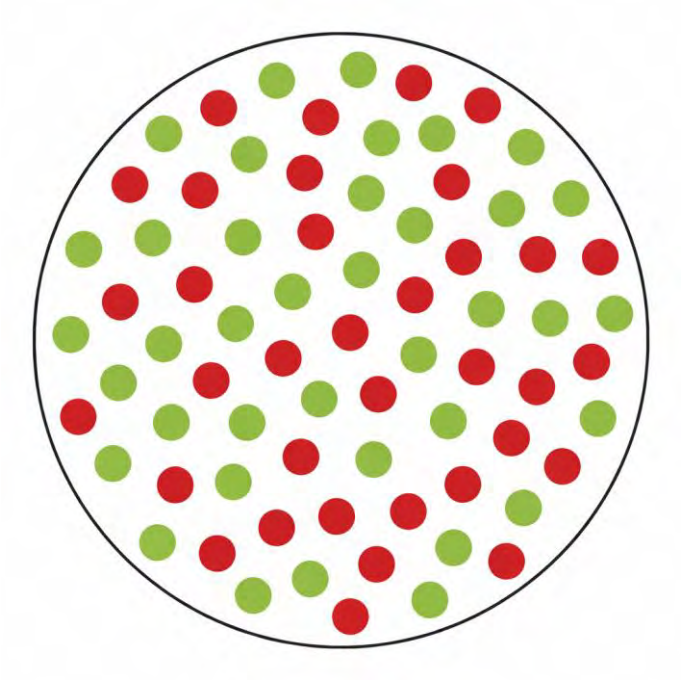


$$f'(x) \approx \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

How to estimate?

- A / B testing
- Switch-back testing

A / B vs Switch-back testing

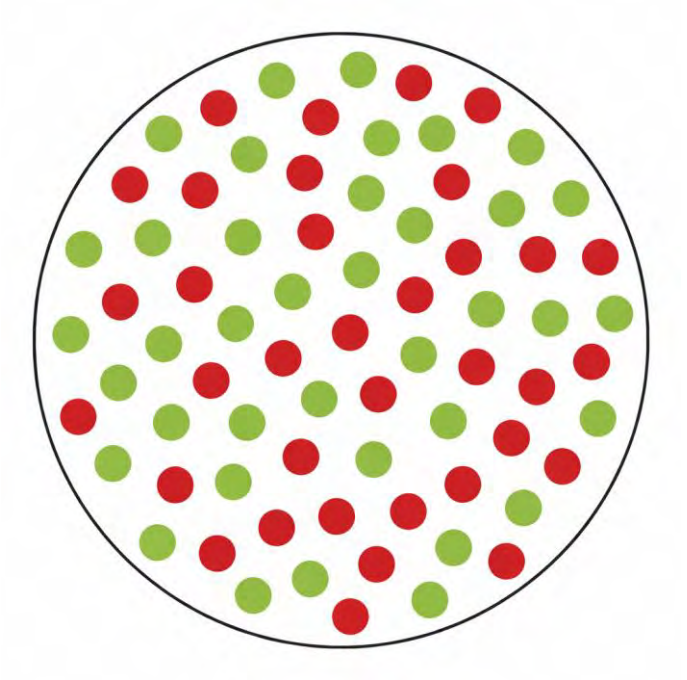


- Simple user randomization
- Affected by user interdependency: e.g. higher delivery fees for one group may lead to excess courier capacity, thus synthetically improving user experience for the other group

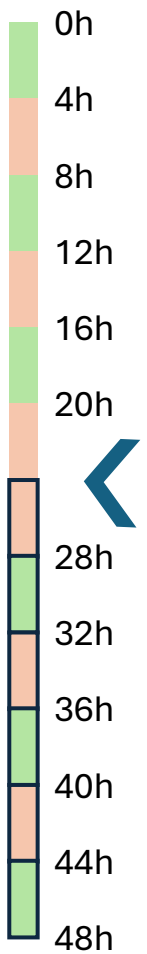
- Two pricing algorithms are used alternately
- All orders are priced according to the same algorithm within each period
- Eliminates interdependency effects between users: e.g. the common pool of couriers

If **window size** chosen is too large – the results of experiments may be affected by periods-related effects (e.g. lunchtime)
If it is too small – the system won't have enough time to react fully

A / B vs Switch-back testing



- Simple user randomization
- Affected by user interdependency: e.g. higher delivery fees for one group may lead to excess courier capacity, thus synthetically improving user experience for the other group



- Two pricing algorithms are used alternately
- All orders are priced according to the same algorithm within each period
- Eliminates interdependency effects between users: e.g. the common pool of couriers

Flip the mode: to eliminate periods-related effects

If **window size** chosen is too large – the results of experiments may be affected by periods-related effects (e.g. lunchtime)
If it is too small – the system won't have enough time to react fully

A / B testing results evaluation: classical statistical tests

E.g. for the probability of a session to end up with an order, the average order size, revenue or profit per order, …

- Two-Sample t-Test (Student's t-test)
- Welch's t-Test (more Robust)
- Mann-Whitney U Test (Wilcoxon Rank-Sum), Permutation Test, etc.

Welch's t-test:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{\Delta}}},$$

where

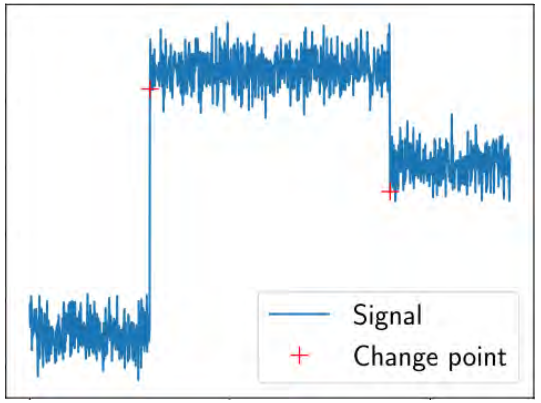
$$s_{\bar{\Delta}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Under the null hypothesis (equal means) the statistic has approximately a Student's t distribution, thus sample statistic value can be compared against this distribution's quantile.

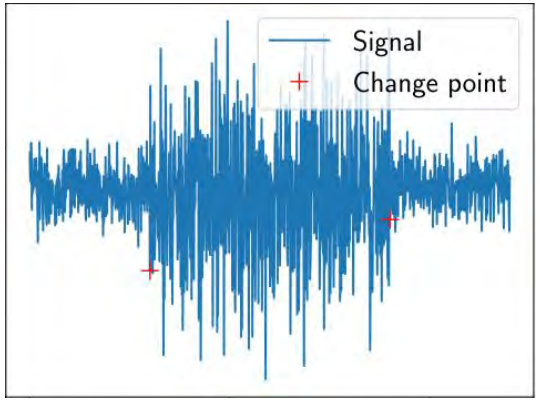
Switch-back testing results evaluation: the change-point problem in time-series analysis

The problem of identifying times when the probability distribution of a stochastic process changes

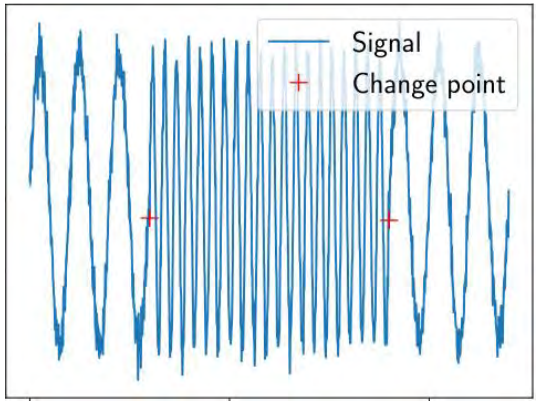
Change in average



Change in variance



Change in frequency

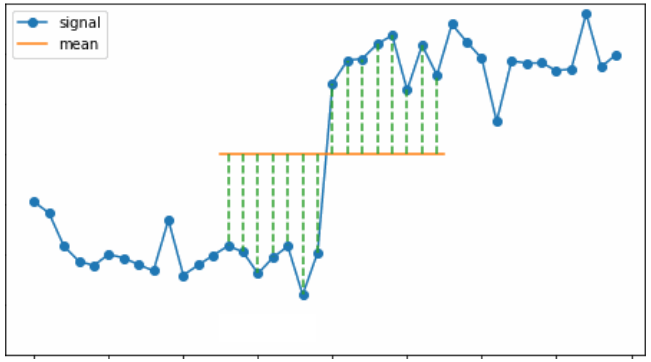


“Yes, we know the moment, when the change is assumed to happen in the metrics we are monitoring – however, we don’t know the size of the jump in the metrics (and whether it happens at all)”

Switch-back testing results evaluation: the change-point problem in time-series analysis

Common approach

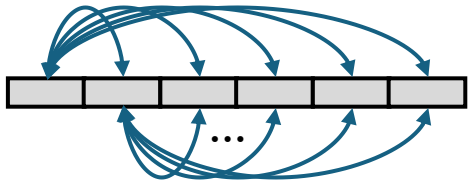
- Is based on maintaining a sliding window of most recent observations
- Mean value is usually computed (exponentially weighted moving average or GARCH model can be used, as well as simple moving average)



Switch-back testing results evaluation: the change-point problem in time-series analysis

An estimate for decision threshold should be produced:

- Choose a recent stable-regime period (without interventions)
- Split it into windows, compute average values per window and compute pairwise absolute differences between average values
- Choose the threshold as a quantile (e.g. 0.95) of the empirical distribution of computed absolute diffs



- the change-points discovered can be validated against switchback experiment periods
- the impact can be estimated as the difference that exceeded the threshold
- the approach can be used to track absence of unexpected negative impacts of the experiment