Getting Uncertainty Right: Reducing Runtime in Airline A/B Tests

AGIFORS Annual Symposium – 2025 | London

Rutger Lit

NOVEMBER 2025

A/B test in a nutshell

The most common type of business experiment is an A/B test comparing two versions of a product

Randomization	Execution & Data collection	Statistical analysis

An A/B test relies on causal inference and statistical inference

To validate a hypothesis, an experiment involves two types of inferences. Both must be present.

Causal inference:

Is this difference really due to causality and nothing else?

→ Proper randomization & minimize interference

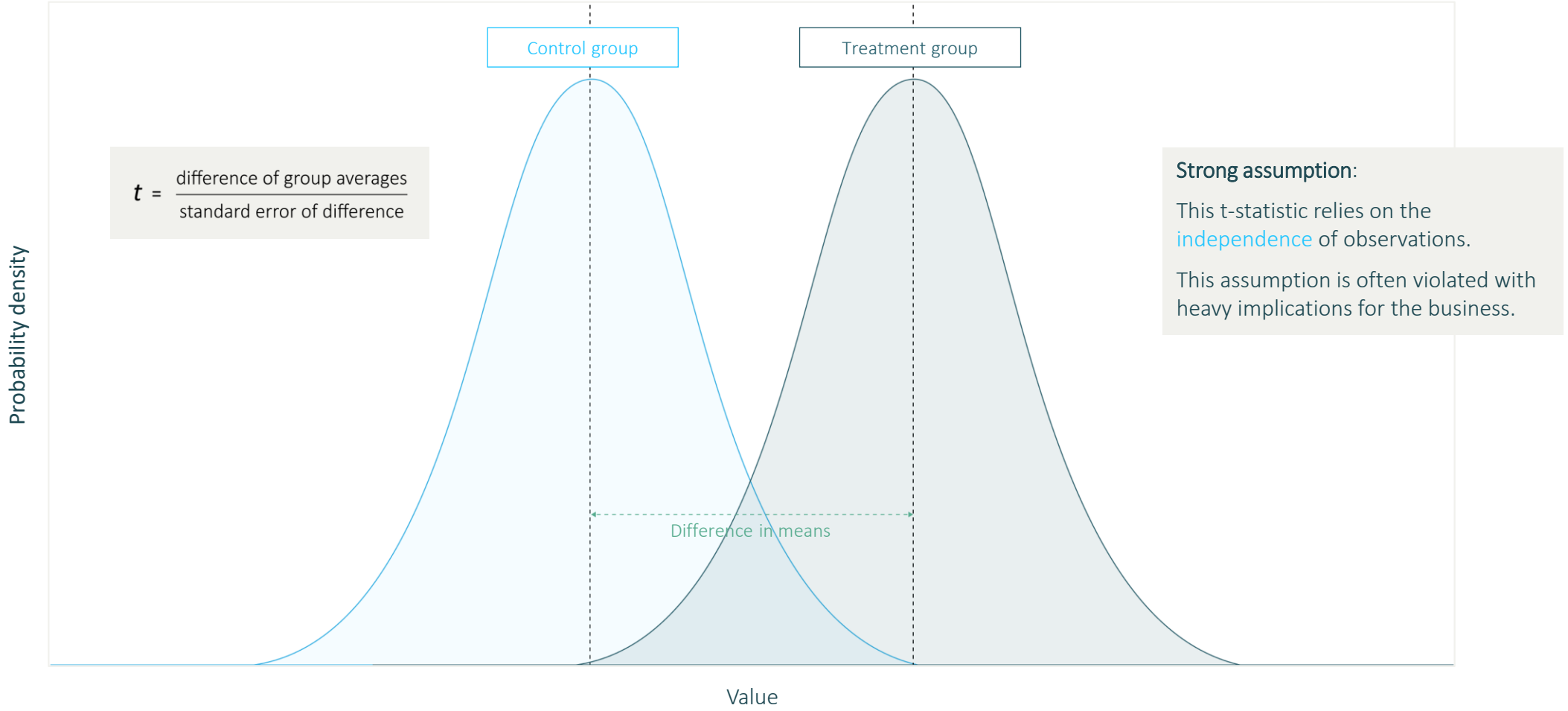
Statistical inference:

Is this difference due to chance?

→ Estimate *statistical uncertainty* to see how likely this difference is due to real signal rather than random noise

Building intuition for uplift measurement

Testing the difference in means



Strong assumption:

This t-statistic relies on the independence of observations.

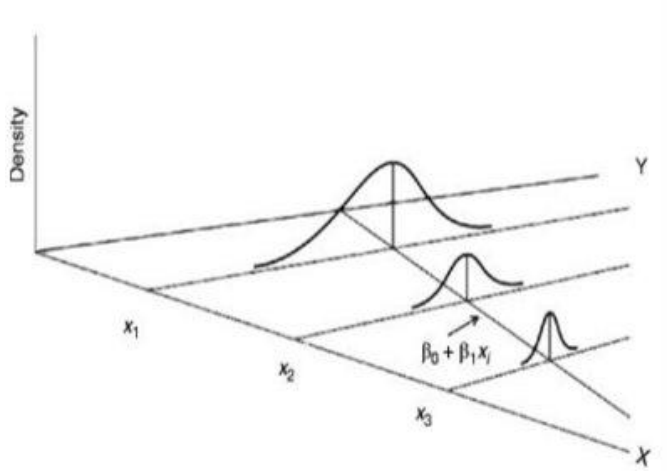
This assumption is often violated with heavy implications for the business.

From independent observations to real-world data

Booking patterns evolve over time and our methods must take that into account

Heteroscedasticity

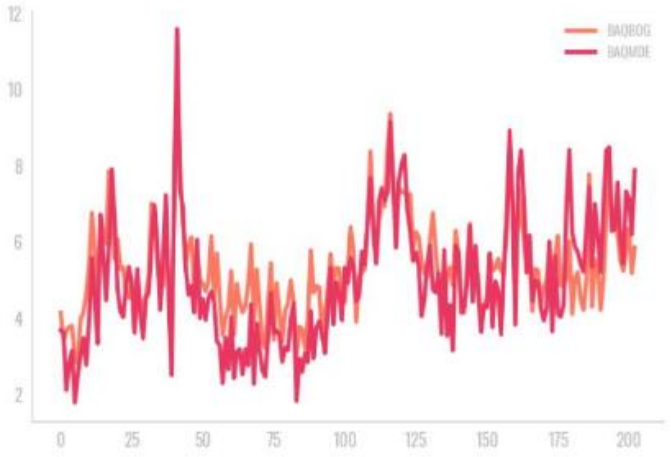
Heterogeneity in measurement error
of different routes



E.g. different level of demand per route

Cross-sectional dependence

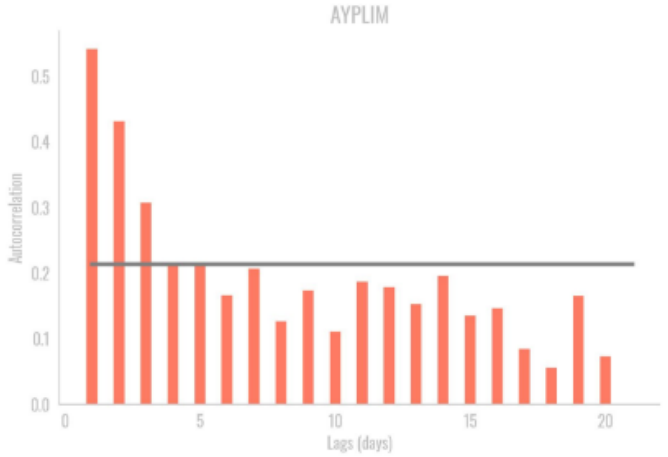
Common effects in rask / revenue
of different routes



E.g. outside changes influence a subset of routes

Time dependence

Auto-correlation in the rask / revenue
of flights within a route



E.g. consistent performance of high revenue routes

Variance under independence

How variance behaves when observations do not influence each other

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

Fundamental variance identity: why autocorrelation increases variance

When observations move together, the covariance term increases total variance.

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + \underbrace{2\mathrm{Cov}(X,Y)}$$

- Cross-sectional dependence
- Time dependence

Our task is to lower the variance of the treatment effect estimate

Two approaches

Starting point

Clustered standard errors

Pros
Correct inference
Simple to implement

Cons
Large standard errors
Long(er) experiment durations

Model the correlation

Generalized Least Squares

Pros
More efficient than OLS
Shorter experiment runtimes

Cons
Requires correct specification
Harder to implement

Adjust experiment design

Switchback design

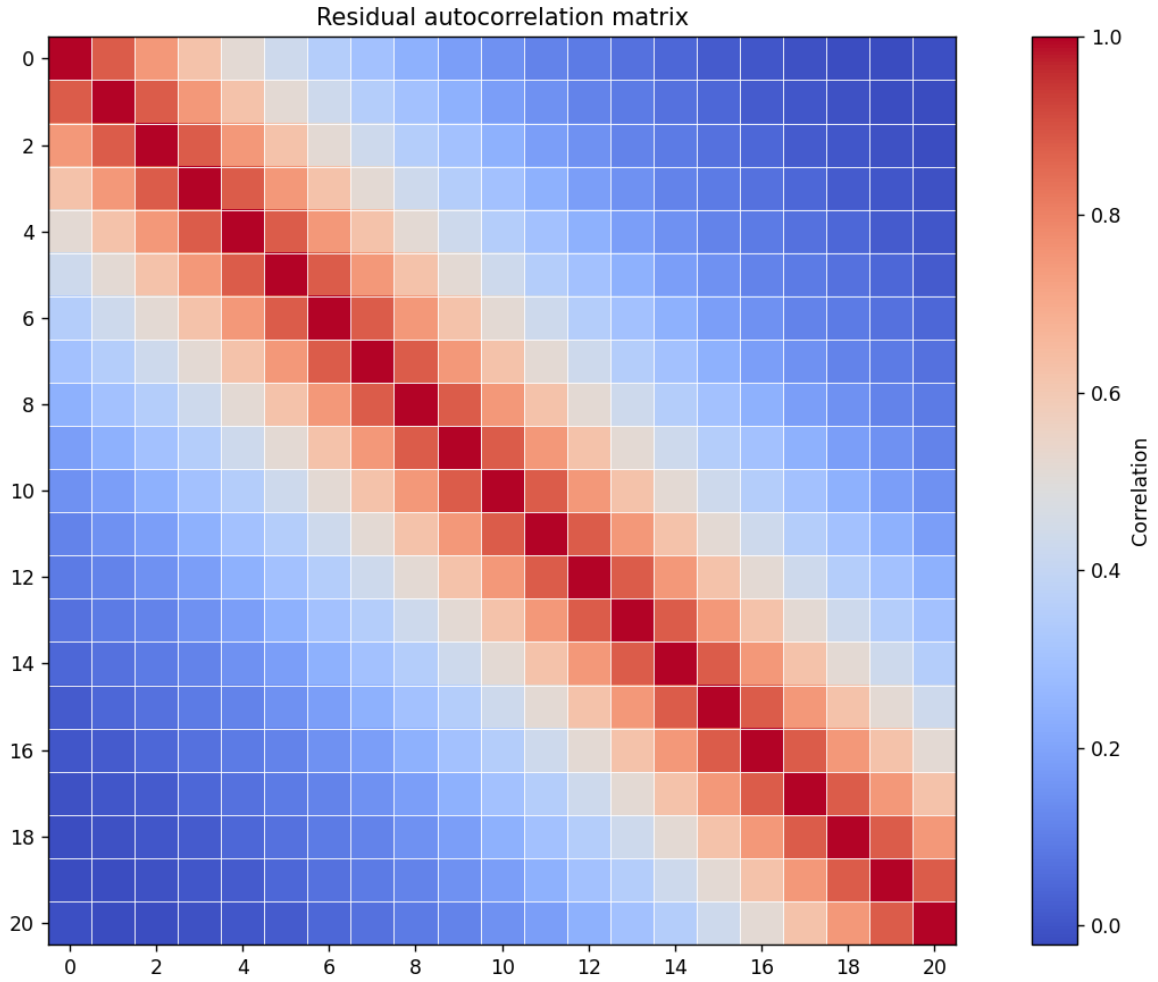
Pros
Mitigates dependencies
Shorter experiment runtimes

Cons
Operational complexity
Interference across periods

Generalized Least Squares

Model the autocorrelation pattern in the residuals

Example residual pattern



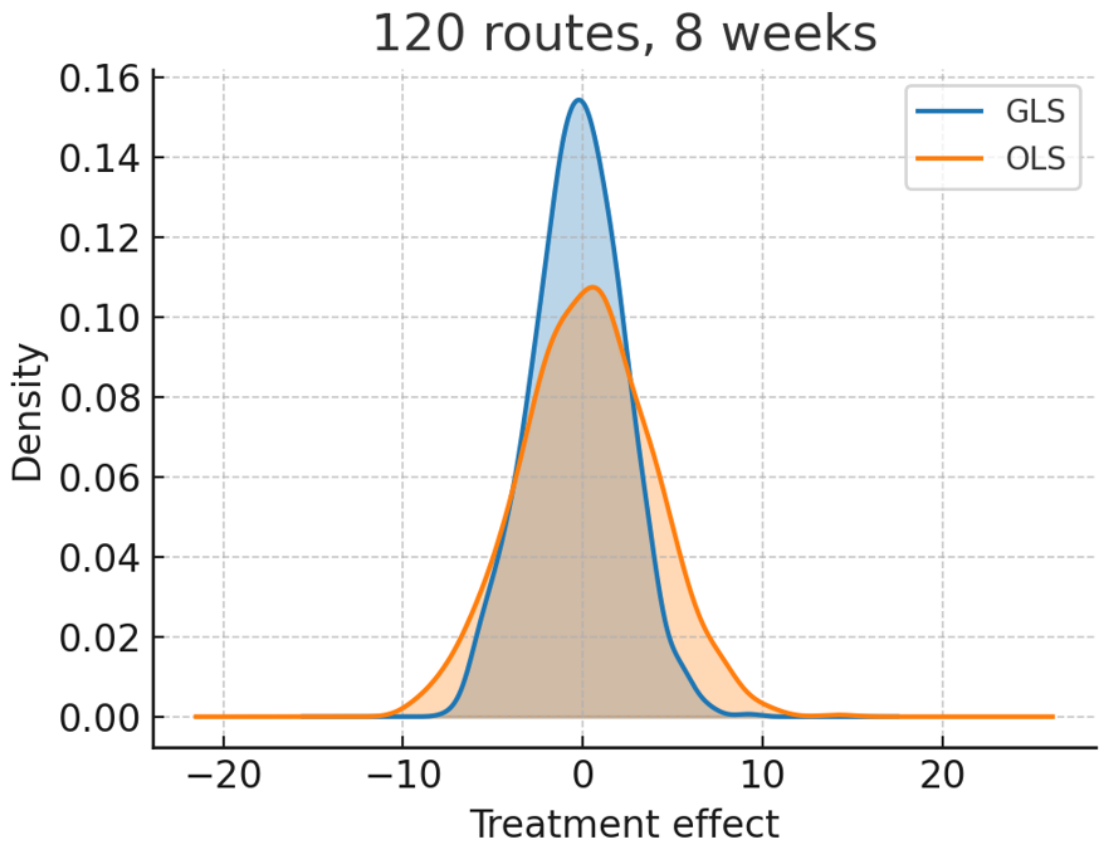
1. Estimating and removing autocorrelation from the data

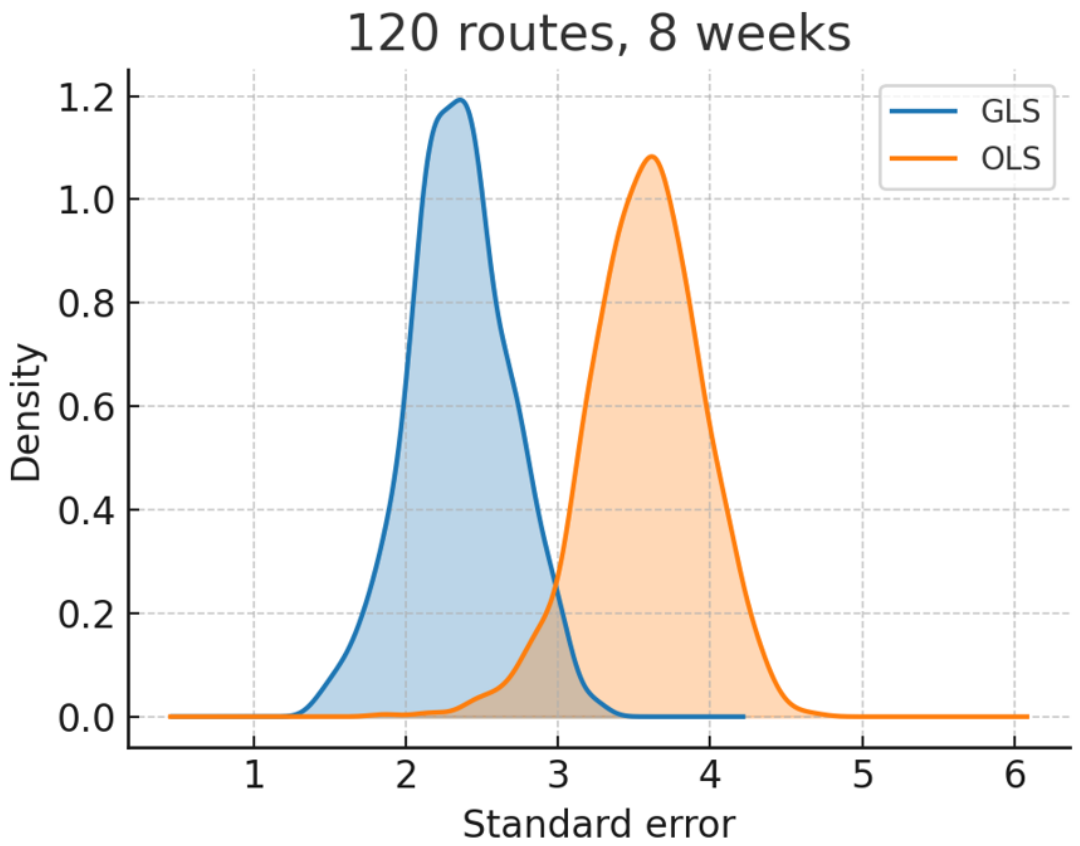
- Run a baseline OLS to obtain residuals
- Use obtained residuals to estimate the autocorrelation structure (e.g. ARMA(1,1))
- Create transformation matrix that removes autocorrelation
- Apply transformation matrix to the data --> transformed data with autocorrelation removed

2. Use transformed data to re-estimate the model

Simulation study

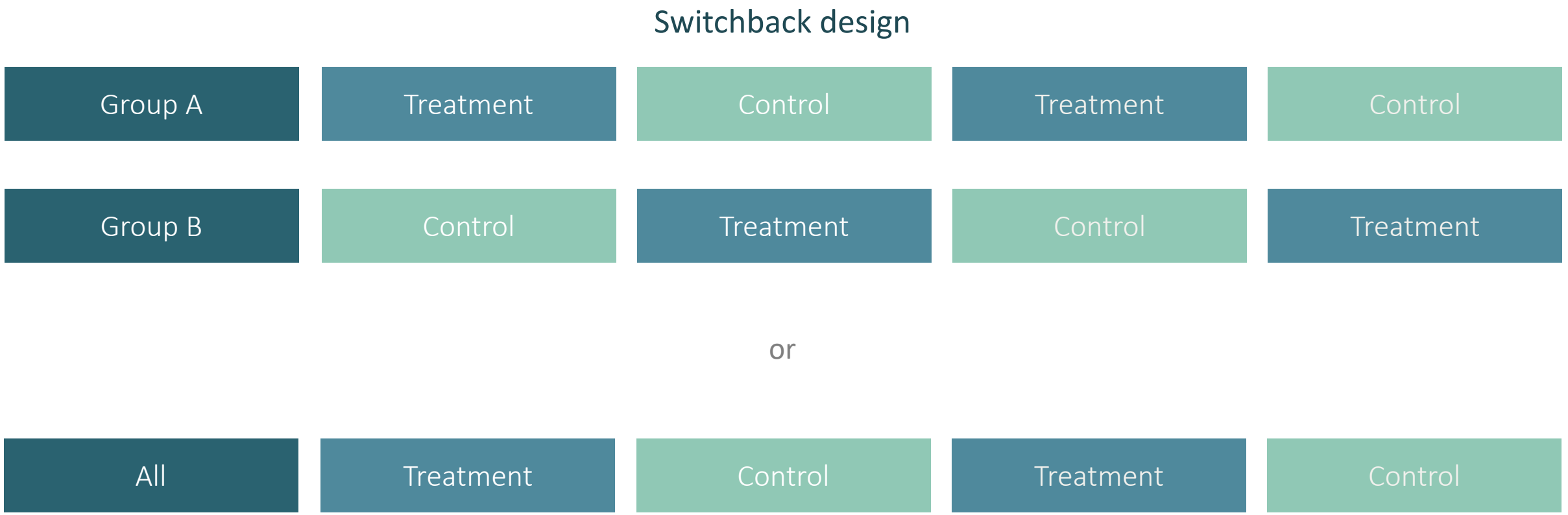
Repeating an A/A experiment 1000 times to measure baseline variation





Switchback Design

There can be various switchback designs to increase efficiency

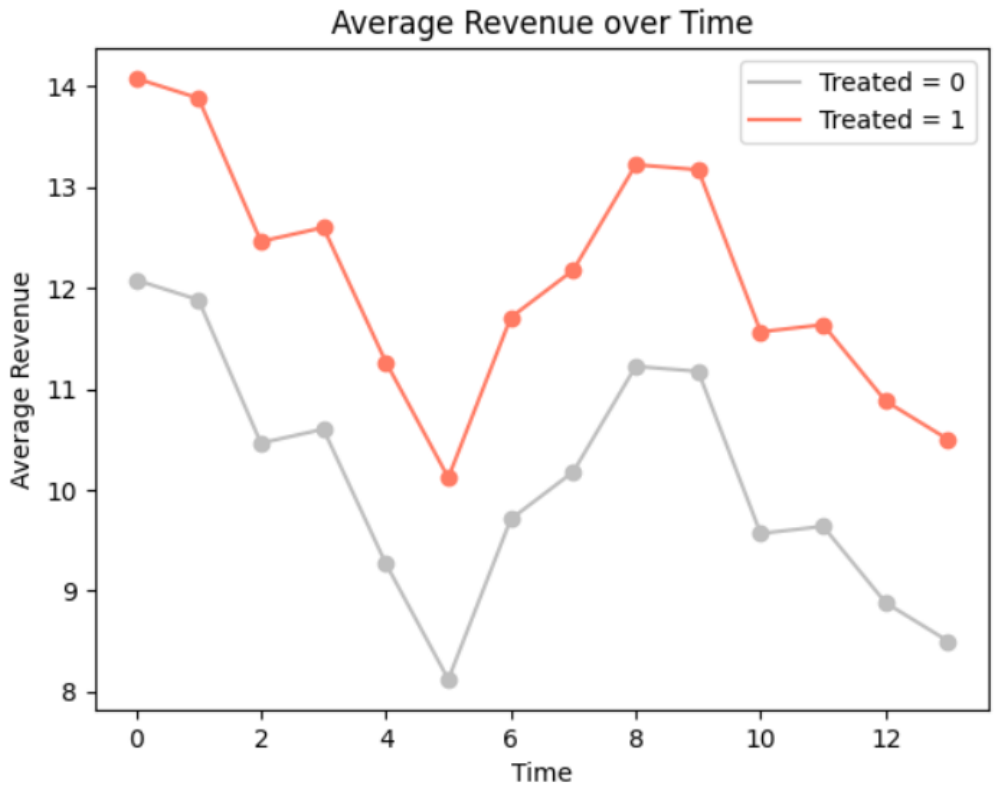


Advantages of using switchback design

Leveraging comparisons across units and periods

Alleviating the influence of autocorrelation

Vanilla A/B testing only allows for the comparisons between groups



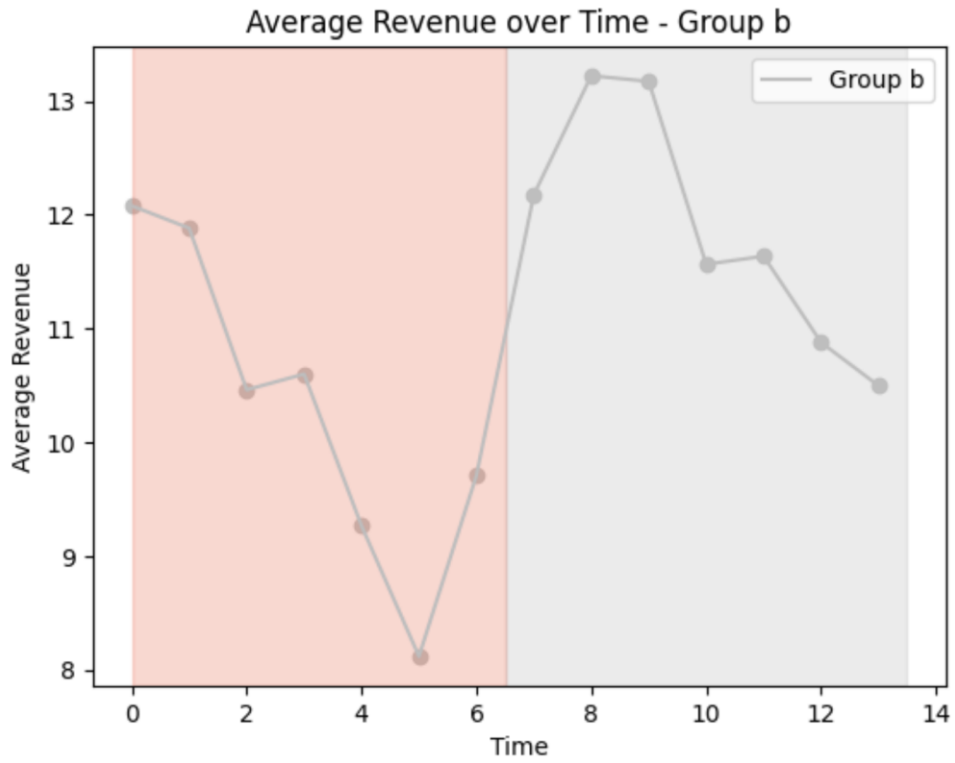
Switchback enables more comparisons

Switchback enables comparisons not only across groups but also overtime which provides additional information on changes within the group given treatment.

Switchback allows for comparison across time

Compare that group's own performance in treated periods versus control periods.

treated control



control treated

Autocorrelation decreases precision

Today's revenue is influenced by past revenue due to ongoing trends, or other factors that persist over time.
Serially correlated data contains less information, which increases the (properly computed) uncertainty.

Switchback design reduces autocorrelation

By comparing the outcomes at different time points within the same group, the underlying similarity across the outcomes (=autocorrelation) can partially be removed.

Lower autocorrelation

Higher precision

More frequent switches, less autocorrelation

Less Autocorrelation

The results of lower variances

Faster, more reliable experiments = more revenue

Better inference improves business performance

Shorter runtimes with valid statistics

→

faster learning cycles

Lower risk of false “wins”

→

scale only what works

More experiments per year

→

more opportunities to grow revenue

Higher confidence in results

→

stronger adoption by stakeholders

Reliable uplift measurement accelerates profitable decision-making.

Broad applicability across airline decisions

Useful across the entire commercial organization

Not just pricing but relevant to every study group.

Fare & pricing experiments

Fuel-saving initiatives

Ancillary revenue

Operational experiments

Marketing & personalization

Any intervention measured with repeated observations benefits from this approach.

Thank you for your attention!