Experimentation, Causal Inference, Metrics, Modeling, and MLOps (Everything Explained)

Diogo Ribeiro

Data Science Lead | Mathematics

November 16, 2025

Agenda

1. A/B Testing: Core Concepts
2. Hypothesis Testing & Power
3. Multiple Testing & Sequential Designs
4. Data Quality & Integrity
5. Interference & Network Effects
6. When A/B Isn't Feasible
7. Modeling & Evaluation
8. Monitoring & Drift
9. MLOps & Delivery
10. Visualization Principles
11. SQL & Warehousing
12. Retention & Survival
13. Pitfalls & Remedies
14. Glossary (Abbreviations)
15. Appendix: Formulas & Checks

Roadmap

1. A/B Testing: Core Concepts
2. Hypothesis Testing & Power
3. Multiple Testing & Sequential Designs
4. Data Quality & Integrity
5. Interference & Network Effects
6. When A/B Isn't Feasible
7. Modeling & Evaluation
8. Monitoring & Drift
9. MLOps & Delivery
10. Visualization Principles
11. SQL & Warehousing
12. Retention & Survival
13. Pitfalls & Remedies
14. Glossary (Abbreviations)
15. Appendix: Formulas & Checks

A/B Testing (Controlled Experiments)

Definition: Randomly split units into Control (A) and Treatment (B) and compare outcomes.

Why: Randomization balances observed/unobserved factors $\Rightarrow$ causal attribution.

Units of Randomization: user, session, cluster/geo.

Key rule: use the smallest unit that avoids *interference*.

- Stable hashing for assignment (e.g., `hash(user_id) mod K`).
- Stratification (blocking) by country/device to reduce variance.
- Exposure integrity: only eligible, actually-exposed users in analysis.

Metrics and Decisioning

- **Primary metric:** single pre-declared decision metric (e.g., D7 retention).
- **Secondary metrics:** additional success indicators (adoption, time-to-value).
- **Guardrails:** safety metrics that must not degrade (e.g., p95 latency, crash-free sessions).
- **KPI:** Key Performance Indicator; connects work to OKRs (Objectives & Key Results).

Roadmap

1. A/B Testing: Core Concepts
2. **Hypothesis Testing & Power**
3. Multiple Testing & Sequential Designs
4. Data Quality & Integrity
5. Interference & Network Effects
6. When A/B Isn't Feasible
7. Modeling & Evaluation
8. Monitoring & Drift
9. MLOps & Delivery
10. Visualization Principles
11. SQL & Warehousing
12. Retention & Survival
13. Pitfalls & Remedies
14. Glossary (Abbreviations)
15. Appendix: Formulas & Checks

Statistical Testing Basics

- **Null (H_0):** no effect; **Alternative (H_1):** effect exists.
- **p-value:** probability of stats as extreme as observed, if H_0 true.
- α **(Type I error):** false positive rate (commonly 0.05).
- β **(Type II error):** false negative rate; **Power** $= 1 - \beta$.
- **Confidence Interval (CI):** Range that would contain the true effect in repeated samples (under the model).

Sample Size & MDE (Minimum Detectable Effect)

Two-proportion sample size per arm (approx.):

$$n \approx \frac{2\,\bar{p}(1-\bar{p})\left(z_{1-\alpha/2}+z_{1-\beta}\right)^2}{\mathrm{MDE}^2}$$

- $\bar{p}$: baseline rate (e.g., 0.12); $z_{1-\alpha/2} \approx 1.96$ for $\alpha = 0.05$; $z_{1-\beta} \approx 0.84$ for 80% power.
- **Duration:** days $\approx \frac{n}{\text{eligible users/day/arm}}$, then round to full weeks to cover seasonality.

Variance Reduction: CUPED

CUPED = Controlled Experiments Using Pre-Experiment Data.

- Use pre-period covariate X correlated with outcome Y.

Adjustment: $Y^* = Y - \theta\,(X - \mathbb{E}[X]), \quad \theta = \frac{\mathrm{Cov}(Y,X)}{\mathrm{Var}(X)}$

Variance factor: $\mathrm{Var}(Y^*) \approx (1 - R^2)\,\mathrm{Var}(Y)$ where R^2 comes from regressing Y on X.

Roadmap

1. A/B Testing: Core Concepts
2. Hypothesis Testing & Power
3. Multiple Testing & Sequential Designs
4. Data Quality & Integrity
5. Interference & Network Effects
6. When A/B Isn't Feasible
7. Modeling & Evaluation
8. Monitoring & Drift
9. MLOps & Delivery
10. Visualization Principles
11. SQL & Warehousing
12. Retention & Survival
13. Pitfalls & Remedies
14. Glossary (Abbreviations)
15. Appendix: Formulas & Checks

Multiple Testing

- Testing many variants/slices inflates false positives.
- **FWER** (Family-Wise Error Rate): Prob(≥ 1 false positive).
- **FDR** (False Discovery Rate): Expected fraction of false among declared positives.
- **Controls:** Holm–Bonferroni (FWER, more powerful than Bonferroni); Benjamini–Hochberg (FDR).

Sequential Testing (Interim Looks)

- **Peeking** inflates Type I error.
- **Fixed-horizon:** Decide once at the end.
- **Alpha-spending:** e.g., O'Brien–Fleming boundaries allocate α across interim looks with conservative early thresholds.

Roadmap

1. A/B Testing: Core Concepts
2. Hypothesis Testing & Power
3. Multiple Testing & Sequential Designs
4. Data Quality & Integrity
5. Interference & Network Effects
6. When A/B Isn't Feasible
7. Modeling & Evaluation
8. Monitoring & Drift
9. MLOps & Delivery
10. Visualization Principles
11. SQL & Warehousing
12. Retention & Survival
13. Pitfalls & Remedies
14. Glossary (Abbreviations)
15. Appendix: Formulas & Checks

SRM (Sample Ratio Mismatch)

Definition: Observed allocation differs from expected (e.g., 50/50 planned, 54/46 observed).

Why it matters: Indicates routing/eligibility/bot issues that can bias estimates.

Detection: χ^2 goodness-of-fit on counts; alert if $p < 0.001$ sustained.

Exposure Integrity & Instrumentation

- Idempotent events with keys; link exposure → outcome.
- Normalize time zones (store in UTC); handle late events with watermarks.
- Exclude bots/internal traffic; audit coverage and eligibility.

Roadmap

1. A/B Testing: Core Concepts
2. Hypothesis Testing & Power
3. Multiple Testing & Sequential Designs
4. Data Quality & Integrity
5. Interference & Network Effects
6. When A/B Isn't Feasible
7. Modeling & Evaluation
8. Monitoring & Drift
9. MLOps & Delivery
10. Visualization Principles
11. SQL & Warehousing
12. Retention & Survival
13. Pitfalls & Remedies
14. Glossary (Abbreviations)
15. Appendix: Formulas & Checks

When Independence Fails

- **Interference:** One unit's treatment affects another's outcome (social features, shared infra).
- **Mitigations:** cluster randomization (geo/store), switchback experiments, measure spillovers.
- Use cluster-robust standard errors in inference.

Roadmap

1. A/B Testing: Core Concepts
2. Hypothesis Testing & Power
3. Multiple Testing & Sequential Designs
4. Data Quality & Integrity
5. Interference & Network Effects
6. When A/B Isn't Feasible
7. Modeling & Evaluation
8. Monitoring & Drift
9. MLOps & Delivery
10. Visualization Principles
11. SQL & Warehousing
12. Retention & Survival
13. Pitfalls & Remedies
14. Glossary (Abbreviations)
15. Appendix: Formulas & Checks

Difference-in-Differences (DiD)

- Compare before/after changes between treated and control groups.
- **Assumption:** Parallel trends.
- **Good practice:** Event-study plots; cluster-robust SEs; wild bootstrap if few clusters.

Synthetic Control, RDD, IV

- **Synthetic Control:** Weighted donor pool mimics treated pre-period; validate via placebo-in-space/time.
- **RDD:** Treatment at threshold; check manipulation (McCrary), estimate locally with optimal bandwidth.
- **IV:** Instrument Z affects treatment T but not outcome Y directly; requires relevance, exogeneity, exclusion.

Propensity Scores (PSM/PSW)

- Model $P(T = 1 \mid X)$ to match/weight units and balance covariates.
- **Diagnostics:** Standardized Mean Difference (SMD) < 0.1, overlap, no extreme weights.
- Sensitivity: Rosenbaum bounds for unobserved confounding.

Roadmap

1. A/B Testing: Core Concepts
2. Hypothesis Testing & Power
3. Multiple Testing & Sequential Designs
4. Data Quality & Integrity
5. Interference & Network Effects
6. When A/B Isn't Feasible
7. Modeling & Evaluation
8. Monitoring & Drift
9. MLOps & Delivery
10. Visualization Principles
11. SQL & Warehousing
12. Retention & Survival
13. Pitfalls & Remedies
14. Glossary (Abbreviations)
15. Appendix: Formulas & Checks

Logit/Probit (Binary Models)

- **Logit:** $\mathrm{logit}(p) = \log \frac{p}{1-p} = \beta_0 + \beta^\top x$; odds ratios e^{β_j}.
- **Probit:** $\Phi^{-1}(p) = \beta_0 + \beta^\top x$; similar to logit.
- **Regularization:** L1 (sparsity), L2 (stability/multicollinearity).
- **Calibration:** Reliability plots; Platt scaling or isotonic regression.

Tree Ensembles & Neural Nets

- **Random Forest:** bagged trees; robust; limited tuning.
- **GBMs (XGBoost/LightGBM):** sequential trees; strong on tabular data.
- **NNs:** FFN (tabular), CNN (images), RNN/LSTM (sequences). Regularize with dropout/weight decay.

Imbalance & Metrics

- **Imbalance:** Class weights/focal loss; avoid SMOTE in time series.
- **Metrics:** ROC-AUC, PR-AUC, Precision, Recall, F1, Brier score (probability calibration).
- **Explainability:** Feature importance, PDP/ICE, SHAP (global & local).

Roadmap

1. A/B Testing: Core Concepts
2. Hypothesis Testing & Power
3. Multiple Testing & Sequential Designs
4. Data Quality & Integrity
5. Interference & Network Effects
6. When A/B Isn't Feasible
7. Modeling & Evaluation
8. **Monitoring & Drift**
9. MLOps & Delivery
10. Visualization Principles
11. SQL & Warehousing
12. Retention & Survival
13. Pitfalls & Remedies
14. Glossary (Abbreviations)
15. Appendix: Formulas & Checks

Drift Types & Tests

- **Data drift:** feature distribution shifts. **Concept drift:** relationship changes.
- **PSI (Population Stability Index):** binned divergence; org thresholds (e.g., 0.1, 0.25).
- **KS test:** max CDF distance; sensitive on large *n*.

Retraining & Ops

- **Triggers:** retrain on drift thresholds or performance decay.
- **Cadence:** scheduled retrains with backtesting before promotion.
- Track schema checks, latency, decision logs.

Roadmap

1. A/B Testing: Core Concepts
2. Hypothesis Testing & Power
3. Multiple Testing & Sequential Designs
4. Data Quality & Integrity
5. Interference & Network Effects
6. When A/B Isn't Feasible
7. Modeling & Evaluation
8. Monitoring & Drift
9. MLOps & Delivery
10. Visualization Principles
11. SQL & Warehousing
12. Retention & Survival
13. Pitfalls & Remedies
14. Glossary (Abbreviations)
15. Appendix: Formulas & Checks

Ops & Tooling

- **CI/CD:** Continuous Integration/Delivery—tests, builds, deploys.
- **Canary:** small % rollout; measure before expand.
- **Shadow:** parallel predictions; no user impact.
- **Feature store:** consistent batch/online features.
- **Model registry:** versions, lineage, approvals.
- **Kill switch:** instant rollback.

Roadmap

1. A/B Testing: Core Concepts
2. Hypothesis Testing & Power
3. Multiple Testing & Sequential Designs
4. Data Quality & Integrity
5. Interference & Network Effects
6. When A/B Isn't Feasible
7. Modeling & Evaluation
8. Monitoring & Drift
9. MLOps & Delivery
10. **Visualization Principles**
11. SQL & Warehousing
12. Retention & Survival
13. Pitfalls & Remedies
14. Glossary (Abbreviations)
15. Appendix: Formulas & Checks

Design for Decision

- Start from the decision/question; one message per slide.
- Maximize data-ink ratio (Tufte); remove chart junk.
- Honest axes; colorblind-safe palettes; add context lines/targets.

Roadmap

1. A/B Testing: Core Concepts
2. Hypothesis Testing & Power
3. Multiple Testing & Sequential Designs
4. Data Quality & Integrity
5. Interference & Network Effects
6. When A/B Isn't Feasible
7. Modeling & Evaluation
8. Monitoring & Drift
9. MLOps & Delivery
10. Visualization Principles
11. **SQL & Warehousing**
12. Retention & Survival
13. Pitfalls & Remedies
14. Glossary (Abbreviations)
15. Appendix: Formulas & Checks

Cohort Retention (SQL Skeleton)

```
WITH installs AS (
  SELECT user_id, MIN(event_ts) AS install_ts
  FROM events
  WHERE event_name = ’install’
  GROUP BY 1
),
activity AS (
  SELECT e.user_id,
         DATE_DIFF(’day’, i.install_ts, e.event_ts) AS dfi
  FROM events e
  JOIN installs i USING (user_id)
  WHERE e.event_name = ’app_open’
    AND e.event_ts BETWEEN i.install_ts
                        AND i.install_ts + INTERVAL ’28␣day’
),
dedup AS (
  SELECT user_id, dfi,
         ROW_NUMBER() OVER (
           PARTITION BY user_id, dfi ORDER BY updated_at DESC
         ) AS rn
  FROM activity
)
SELECT dfi,
       COUNT(DISTINCT CASE WHEN dfi = 0 THEN user_id END) AS n0,
       COUNT(DISTINCT CASE WHEN rn = 1 THEN user_id END) AS active,
       COUNT(DISTINCT CASE WHEN rn = 1 THEN user_id END)::float
       / NULLIF(COUNT(DISTINCT CASE WHEN dfi = 0 THEN user_id END), 0)
       AS retention
FROM dedup
GROUP BY 1
ORDER BY 1;
```

Roadmap

1. A/B Testing: Core Concepts
2. Hypothesis Testing & Power
3. Multiple Testing & Sequential Designs
4. Data Quality & Integrity
5. Interference & Network Effects
6. When A/B Isn't Feasible
7. Modeling & Evaluation
8. Monitoring & Drift
9. MLOps & Delivery
10. Visualization Principles
11. SQL & Warehousing
12. **Retention & Survival**
13. Pitfalls & Remedies
14. Glossary (Abbreviations)
15. Appendix: Formulas & Checks

Kaplan–Meier & Cox PH

- **KM:** Nonparametric survival $S(t)$ with censoring.
- **Cox PH:** Hazard model with multiplicative covariate effects; test proportional hazards via Schoenfeld residuals.
- Compare cohorts with log-rank test.

Roadmap

1. A/B Testing: Core Concepts
2. Hypothesis Testing & Power
3. Multiple Testing & Sequential Designs
4. Data Quality & Integrity
5. Interference & Network Effects
6. When A/B Isn't Feasible
7. Modeling & Evaluation
8. Monitoring & Drift
9. MLOps & Delivery
10. Visualization Principles
11. SQL & Warehousing
12. Retention & Survival
13. **Pitfalls & Remedies**
14. Glossary (Abbreviations)
15. Appendix: Formulas & Checks

Common Pitfalls

- Peeking without correction $\Rightarrow$ inflated Type I error.
- Ignoring SRM $\Rightarrow$ biased estimates.
- Leakage from future/post-treatment variables.
- Metric misalignment (optimize clicks vs. retention).
- Multiple testing without correction; Simpson's paradox.
- No post-ship monitoring; regression to the mean.

Roadmap

1. A/B Testing: Core Concepts
2. Hypothesis Testing & Power
3. Multiple Testing & Sequential Designs
4. Data Quality & Integrity
5. Interference & Network Effects
6. When A/B Isn't Feasible
7. Modeling & Evaluation
8. Monitoring & Drift
9. MLOps & Delivery
10. Visualization Principles
11. SQL & Warehousing
12. Retention & Survival
13. Pitfalls & Remedies
14. **Glossary (Abbreviations)**
15. Appendix: Formulas & Checks

Quick Reference

A/B	Control vs. treatment experiment
KPI	Key Performance Indicator
SRM	Sample Ratio Mismatch
MDE	Minimum Detectable Effect
CUPED	Variance reduction using pre-period covariates
FWER	Family-Wise Error Rate
FDR	False Discovery Rate
RCT	Randomized Controlled Trial
DiD	Difference-in-Differences
RDD	Regression Discontinuity Design
IV	Instrumental Variables
PSM/PSW	Propensity Score Matching/Weighting
ROC-AUC/PR-AUC	Discrimination summaries under class imbalance
Brier	Probability calibration error (MSE of probs)
PSI	Population Stability Index
KS	Kolmogorov–Smirnov test
p95	95th percentile (latency tail)
PDP/ICE	Partial Dependence / Individual Conditional Expectation
SHAP	Shapley-based local/global explanations

Roadmap

1. A/B Testing: Core Concepts
2. Hypothesis Testing & Power
3. Multiple Testing & Sequential Designs
4. Data Quality & Integrity
5. Interference & Network Effects
6. When A/B Isn't Feasible
7. Modeling & Evaluation
8. Monitoring & Drift
9. MLOps & Delivery
10. Visualization Principles
11. SQL & Warehousing
12. Retention & Survival
13. Pitfalls & Remedies
14. Glossary (Abbreviations)
15. Appendix: Formulas & Checks

Formulas

- **Two-proportion MDE (given n):** $\mathrm{MDE} \approx \sqrt{\dfrac{2\,\bar{p}(1-\bar{p})\,(z_{1-\alpha/2}+z_{1-\beta})^2}{n}}$
- **CUPED:** $Y^* = Y - \theta(X - \mathbb{E}[X]),\ \theta = \frac{\mathrm{Cov}(Y,X)}{\mathrm{Var}(X)}$
- **BH-FDR:** sort p-values $p_{(i)}$, find largest k with $p_{(k)} \le \frac{k}{m}q$, declare $1..k$.
- **Holm:** order p-values; compare $p_{(i)} \le \frac{\alpha}{m-i+1}$ sequentially.
- **O'Brien–Fleming (alpha-spending):** conservative early, liberal late boundaries.

Checks & Runbooks

- Pre-register primary metric, guardrails, decision rule.
- Powering with realistic MDE; round duration to full cycles.
- Instrumentation dry run; SRM alarms; exposure audits.
- Sensitivity: heterogeneity, alternative tests (Welch/MWU), outliers.
- Post-ship: DiD vs. non-adopters; kill switch; rollback plan.

Closing

Questions & Discussion