Statistical Approaches in Online Testing

Dmytro Skorokhodov

AI Ukraine 17
Kharkiv
Sep 23-24, 2017

Oracle Maxymiser

ORACLE | maxymiser is leading provider of cloud-based software that enables marketers to test, target and personalize what a customer sees on a Web page or mobile app, substantially increasing engagement and revenue

- ❑ **2006** – Foundation
- ❑ **2015** – Acquisition by Oracle

Agenda

1. **Introduction to testing**
    - Testing: When? Where?
    - Testing: Collect evidence
    - Testing: Compare performance
    - Statistical testing

2. **Statistical approaches in testing**
    - Frequentist approach
    - Bayesian approach

3. **Challenges in online testing**
    - How long to run a test?
    - Continuous monitoring
    - Delayed responses
    - 2+ alternatives
    - Throttling
    - Multiple goals
    - Other challenges

Introduction to testing

Testing: When? Where?

Default | Alternative | Change? | TEST

- Result of research
- Breakthrough idea
- Necessary change
- …

- Collect evidence
- Compare performance







ORACLE

Testing: Collect evidence

Define test KPI's

Target audience

- **Visitors**, sessions, views, …

Target metric

- **Clicks**, Purchases, Sign-Ups, …

Success measure

- **Conversion rate**, average revenue, …

Collect evidence

Split target audience

- deterministic
- 'random'
- stratified
- …

Default OR Alternative

- Visitor = trial

Observe target metric

- Click / no click = trial outcome

Testing: Compare performance

Evidence

Variant	Visitors	Clicks*	Conv rate
Default	98	31	31.63%
Alternative	103	34	33.01%

Does evidence tell that Alternative is better than Default?

Given:
Default true conv rate = **30%**
Alternative true conv rate = **29%**

There is
43.74% chances that Alternative will have higher sample conversion rate

Sample estimates

- ≠ true conversion rates

Statistical error

- Sampling error
- Random nature of visitor response
- Imperfect knowledge of future
- Wrong model of experiment

Use statistical test!

Assumption: *0 or 1 click per visitor*

ORACLE

Statistical testing

Formulate hypotheses

Collect evidence

Check hypotheses

- $\boldsymbol{H_0}$ *(null hypothesis)*
  - Equal conversion rates
- $\boldsymbol{H_1}$ *(alternative hypothesis)*
  - Unequal conversion rates

- Serve visitors with Default or Alternative
- Observe clicks

- Perform statistical test
- Declare Alternative as a winner or not

No 100% guarantee that the winner is found

Statistical approaches in testing

Statistical approaches

	Frequentist inference	Bayesian inference
Parameters	Fixed (may be unknown)	Random, may be presented as beliefs
Assumptions	H_0 is true by default	H_0 and H_1 have some prior probabilities
Thresholds	Significance level α	
Evidence (E)	Used to disprove H_0:	Used to update beliefs in H_0 and H_1:
Result	p-value – probability of results to be at least as extreme as evidence given H_0	Calculate posterior probabilities of H_0 and H_1
	Reject H_0 if p-value $< \alpha$, and accept H_0 otherwise	Reject H_0 if $P(H_0 \mid E) < \alpha$, and accept H_0 otherwise

Frequentist approach

Null hypothesis H_0

- $p_D = p_A$, $p_D > p_A$, …

Alternative hypothesis H_1

- $p_D \neq p_A$, $p_D < p_A$, …

Significance level α

- 0.05, 0.01, 0.1, …

Statistical test

- **T-test**, χ^2-test, U-test, …
- applicable to wide family of distributions
- motivated by the law of large numbers

Notations:

- n – number of visitors
- c – number of clicks
- p – true conversion rate
- $\widehat{p}$ – sample conversion rate

T-test details (Two tailed two samples Welch T-test)

Step	Formula
Calculate sample estimates	$\widehat{p}_D = \frac{c_D}{n_D}$ and $\widehat{p}_A = \frac{c_A}{n_A}$
Calculate T-statistics	$t = \dfrac{\lvert \widehat{p}_D - \widehat{p}_A \rvert}{\sqrt{\frac{\widehat{p}_D \cdot (1 - \widehat{p}_D)}{n_D} + \frac{\widehat{p}_A \cdot (1 - \widehat{p}_A)}{n_A}}}$
Calculate P-value	p-value $= 2 \int_t^{+\infty} \varphi(t)\, dt$, φ is standard normal p.d.f.

Frequentist approach: Example

Variant	Visitors	Clicks	Conv rate	T-statistics	P-value
Default	98	31	31.63%	0.209	0.83
Alternative	103	34	33.01%		

Accept H_0 at 0.05 significance level:

- Not enough data to prove that Alternative is different from Default

Variant	Visitors	Clicks	Conv rate	T-statistics	P-value
Default	98	**15**	**15.31%**	3.005	0.027
Alternative	103	34	33.01%		

Reject H_0 at 0.05 significance level:

- Alternative is different from Default with 5% significance

Bayesian approach: one simple coin example

Null hypothesis H_0

- $p = q_1$ with $\pi_0 = P(H_0) = 0.5$ prior probability

Alternative hypothesis H_1

- $p \neq q_2$ with $\pi_1 = P(H_1) = 0.5$ prior probability

Significance level α

- 0.05, 0.01, 0.1, …

Update rule ingredients

- Coin model: p is the success rate
- Bayes theorem: $P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$
- Law of total probability: $P(B) = \sum_j P(B \mid A_j) \cdot P(A_j)$

Update rule details given a heads and b tails

Step	Formula
Posterior probability for H_i	$P(H_i \mid E) = \frac{P(E \mid H_i)}{P(E)} \cdot \pi_i$
Probability of evidence given H_i	$P(E \mid H_i) = q_i^a (1 - q_i)^b$
Probability of evidence	$P(E) = \sum_{i=0}^{1} P(E \mid H_i) \cdot \pi_i$

EXAMPLE	
Assumption: $q_1 = 0.5, q_2 = 0.3$ **Evidence:** 4 heads, 5 tails	$P(E \mid H_0) = 0.5^4 \cdot 0.5^5 \approx 0.002$ $P(E \mid H_1) = 0.3^4 \cdot 0.7^5 \approx 0.0014$ $P(E) \approx 0.0033$ $P(H_0 \mid E) \approx 58.93\%$

NOTE: If $\pi_0 = 0.9$ and $\pi_1 = 0.1$ then $P(H_0 \mid E) \approx 92.81\%$

ORACLE

Bayesian approach: two coins example

Null hypothesis H_0

- $p_D = p_A$ with π_0 prior and $\pi_0(p_D, p_A)$ prior p.d.f. of parameters

Alternative hypothesis H_1

- $p_D \neq p_A$ with π_1 prior and $\pi_1(p_D, p_A)$ prior p.d.f. of parameters

Significance level α

- 0.05

Update rule ingredients

- Coin model: p_D and p_A are the success rates
- Bayes theorem
- Law of total probability
- Bayes theorem for p.d.f's:

$$P(A \mid B) = \frac{P(A)}{P(B)} \int_{\Omega} p(\omega \mid A) d\omega$$

Update rule details given a_j heads and b_j tails for j^{th} coin

Step	Formula
Posterior probability for H_i	$P(H_i \mid E) = \frac{P(E \mid H_i)}{P(E)} \cdot \pi_i$
Posterior p.d.f. of parameters in H_i	$\pi_i(p, q \mid E) = p^{a_0}(1-p)^{b_0} q^{a_1}(1-q)^{b_1} \pi_i(p, q)$
Probability of evidence given H_i	$P(E \mid H_i) = \int_0^1 \int_0^1 \boxed{1} \cdot \pi_i(p, q \mid E)\, dp dq$
Probability of evidence	$P(E) = \sum_{i=0}^{1} P(E \mid H_i) \cdot \pi_i$

$L(p, q)$

$L(p, q) = 1$
$L(p, q) = \max\{q - p; 0\}$
$L(p, q) = |q - p|$
…

Loss function!

Non-comprehensive comparison

	Frequentist inference	Bayesian inference
+	• Simple “universal” indicator • Directly verifiable (AA/AB tests)	• Flexible • Loss function • Optional stopping out of box
−	• No rejection for $\boldsymbol{H_1}$ • $\boldsymbol{p}$**-value** is prone to misinterpretations	• “Subjective” • Difficult to interpretation for a non Statistician • No standard choice for priors, hypotheses, data models • Revenue testing is much more advanced

ORACLE

Challenges in Online Testing

Fixed sample methodology to online testing

Define:
significance level $\boldsymbol{\alpha}$
sample size $\boldsymbol{n}$

Collect data

Stop the test when sample size is reached

Reject H_0
if $\boldsymbol{p}$**-value** $< \boldsymbol{\alpha}$

Accept H_0
if $\boldsymbol{p}$**-value** $> \boldsymbol{\alpha}$

	$\boldsymbol{H_0}$ **is rejected**	$\boldsymbol{H_0}$ **is accepted**
$\boldsymbol{H_0}$ is true	**Type I error**	Correct inference
$\boldsymbol{H_1}$ is true	Correct inference	**Type II error**

Type I error is bounded by $\boldsymbol{\alpha}$

- Ensured by methodology

Type II error has no sense with $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$

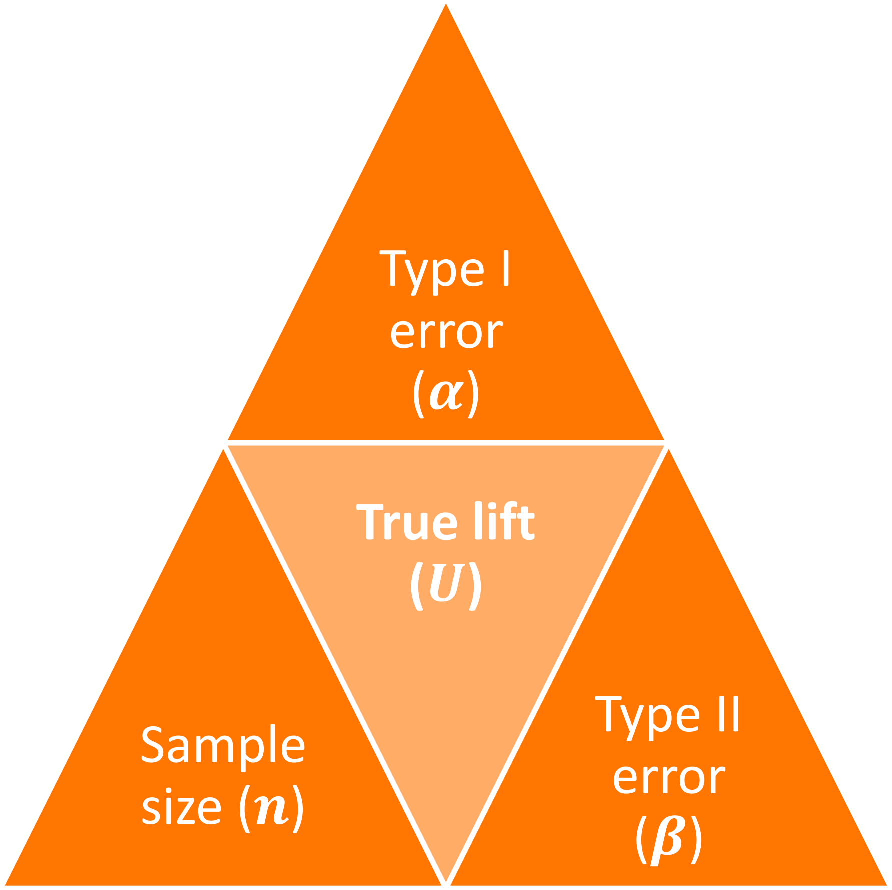
- No distance between hypotheses

Consider different alternative: $H_U : |p_D - p_A| > U$

- Can assign $\boldsymbol{\beta}$ threshold for not rejecting $\boldsymbol{H_0}$

ORACLE

Challenge 1: How long to run a test?



$$n = \frac{\left(\Phi(\beta) + \Phi\left(1 - \frac{\alpha}{2}\right)\right)^2}{U^2 \cdot p}$$

Type I error is bounded by α

- Ensured by methodology

Type II error is bounded by β for H_U

- Ensured by formula

U is pure guess

Challenge 2: Continuous monitoring

Define:
significance level $\boldsymbol{\alpha}$
sample size $\boldsymbol{n}$

Collect data

Stop the test when sample size is reached

Reject H_0 if **p-value** $< \boldsymbol{\alpha}$

Accept H_0 if **p-value** $> \boldsymbol{\alpha}$

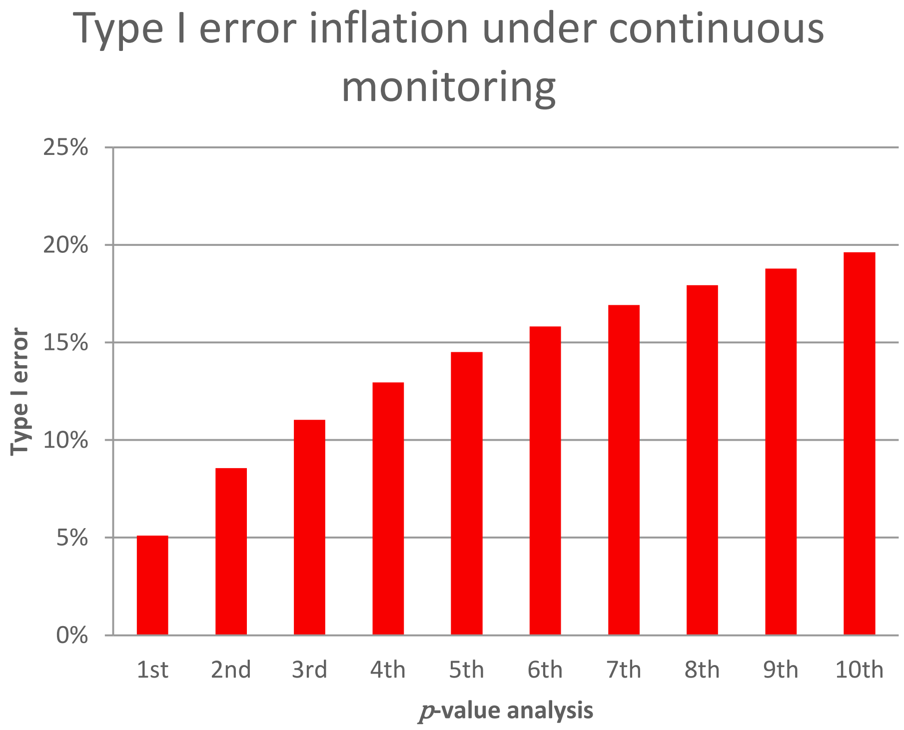
Reject H_0 and stop the test if **p-value** $< \boldsymbol{\alpha}$

Continue the test if **p-value** $> \boldsymbol{\alpha}$

Business need:
(1) Data available in reports asap
(2) Allocate additional time to get results

Continuous monitoring:
Inflated type I error control

Challenge 2: Continuous monitoring inflates type I error



H_0 will be rejected eventually with continuous monitoring!

Law of iterated logarithm

$$\overline{\lim_{k\to\infty}} \frac{T_k}{\sqrt{\log\log k}} = \sqrt{2}, \text{ a.s.}$$

Design a methodology that accounts for continuous monitoring

ORACLE

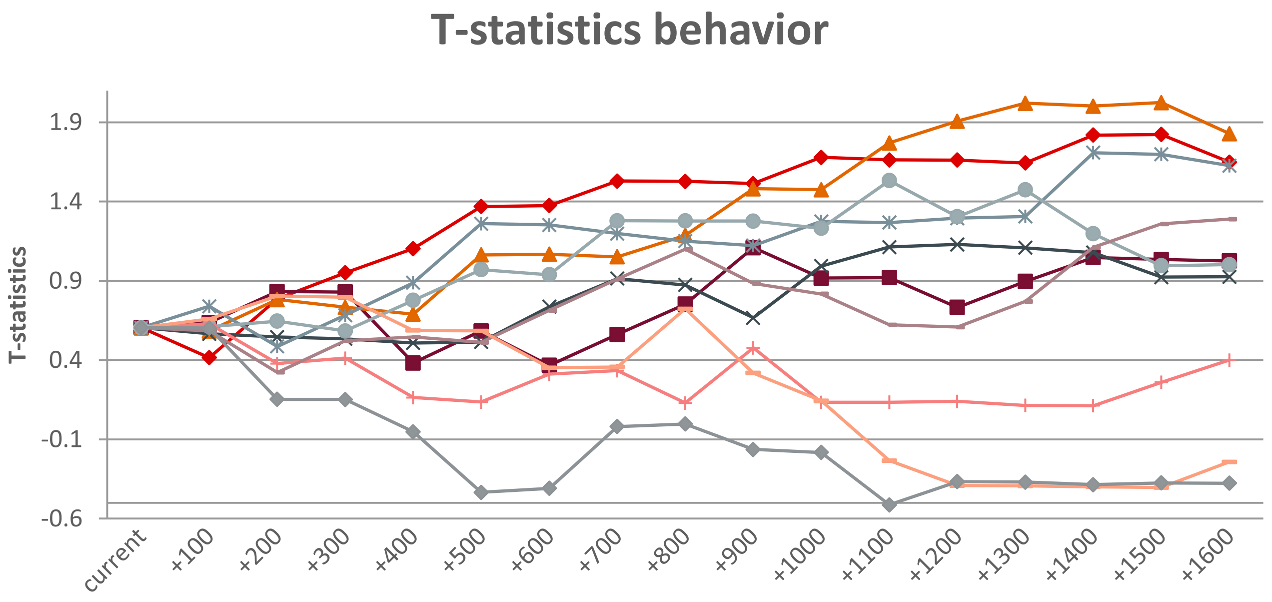
Do sequential testing!

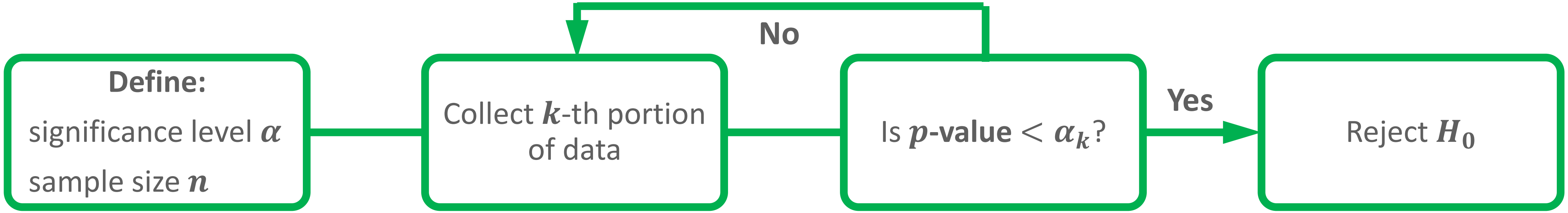
Appeared in 1920's

- A. Wald, J. Wolfowitz, W. Allen Wallis, M. Friedman, H. Robbins, …

Key idea

- **Control type I error** by:
  - Using smaller significance levels – α_k – at interim analyses: $\sum_k \alpha_k < \alpha$
- **Achieve high power** by:
  - Using covariates, *i.e.* similarity between p-**value**'s at consecutive analyses





Challenge 3: Delayed responses & sequential tests

Delayed responses examples

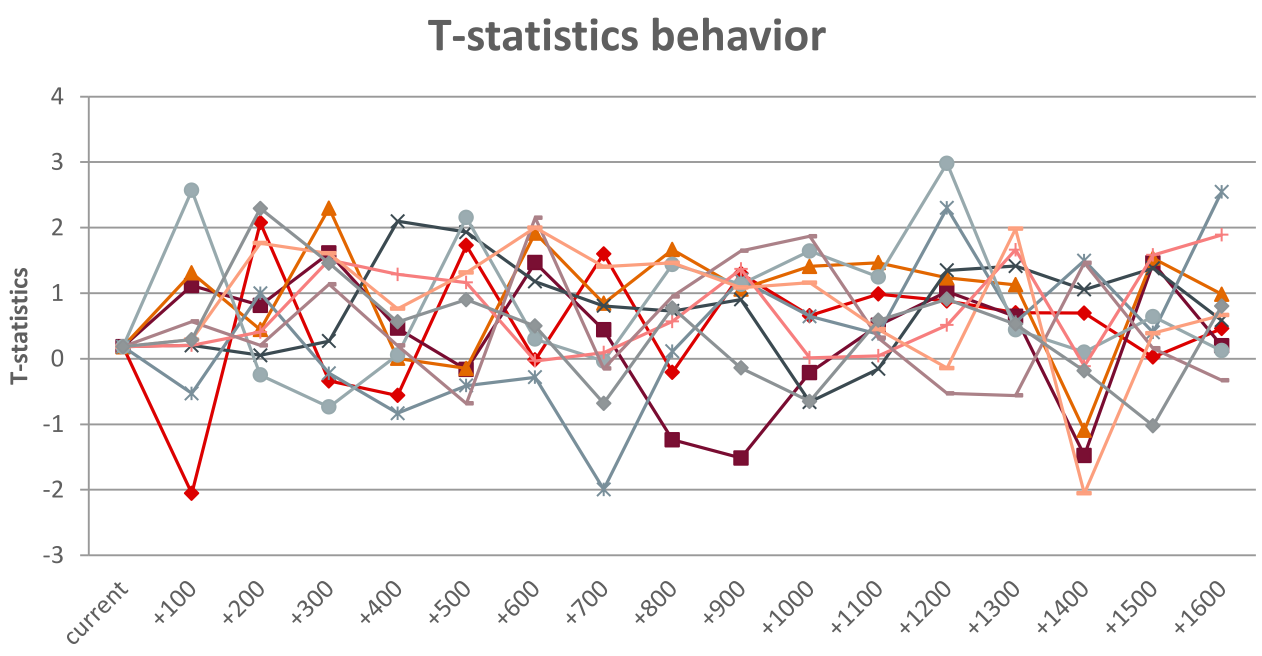
- Purchases, multiple conversions, …

Delayed responses effects

- Previous conclusions may change
- Inflated type I error in sequential tests due to covariance accounting

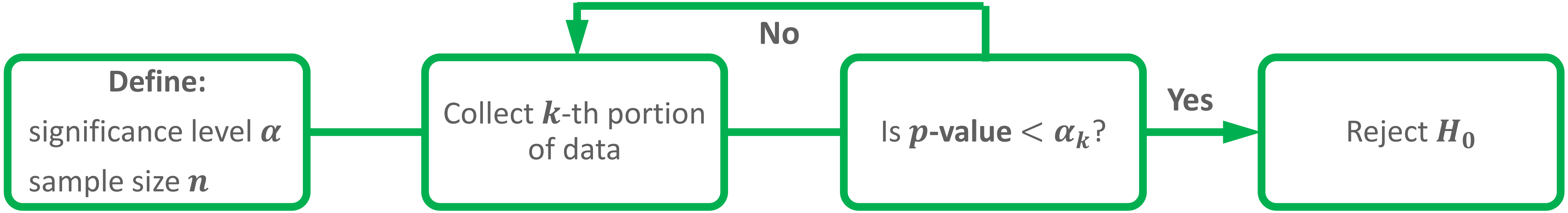
Unknowns with delayed responses

- Percentage of delayed actions
- Distribution of delay



Ignore covariates!

Our approach



Strict control over type I error

- α percent of false positive results

Zero type II error

- Every test with non-zero difference will be concluded eventually

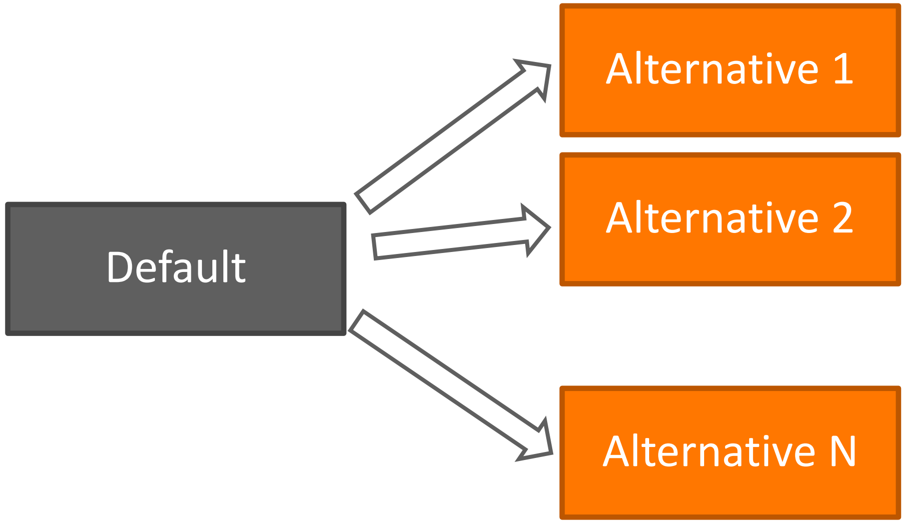
'Flat' test notification

- Receive message if difference is low enough (lower than user-input threshold)

Standard/low traffic modifications

ORACLE

Challenge 4: 2+ alternatives (ABn and MVT tests)



Which is better than Default?

- Formulate multiple (N) null hypothesis:

$$H_{0,1}: p_D = p_{A_1}, \quad H_{0,2}: p_D = p_{A_2}, \quad \ldots, \quad H_{0,N}: p_D = p_{A_N}$$

- Protect against type I error inflation:

Family-wise error – probability of rejecting 1+ true null hypothesis

Bonferroni – multiply individual p**-values** by N
Holm-Bonferroni

False discovery rate – expected proportion of incorrectly rejected null hypotheses among all rejected null hypotheses

Benjamini-Hochberg

Challenge 4: 2+ alternatives (ABn and MVT tests)

More data is needed to reach statistical significance

- Exclude bad performing variants (ABn & MVT)
- Neglect some degree of factors interaction (MVT)
  - Orthogonal arrays
  - Taguchi
  - Fractional factorial designs
  - Optimal designs

ORACLE

Challenge 5: Throttling mid-test

Business need:
Validate Alternative on small portion of traffic and increase this proportion later if it proves competitive against the Default

Example

	Default			Alternative		
	Visitors	Converters	Conv Rate	Visitors	Converters	Conv Rate
1st week	9000	900	10.0%	1000	105	10.5%
2nd week	5000	450	9.0%	5000	455	9.1%
Total	**14000**	**1350**	**9.6%**	**6000**	**560**	**9.3%**

Problem (known as **Simpson's paradox):**
Standard (cumulative) estimates are skewed!

ORACLE

Challenge 5: Throttling mid-test

Cumulative estimate

- $\frac{c_1+c_2+\cdots+c_n}{v_1+v_2+\cdots+v_n}$

Inverse probability weighting

- $\frac{\frac{c_1}{\pi_1}+\frac{c_2}{\pi_2}+\cdots+\frac{c_n}{\pi_n}}{\frac{v_1}{\pi_1}+\frac{v_2}{\pi_2}+\cdots+\frac{v_n}{\pi_n}}$

v_j is the number of visitors on period j
c_j is the number of converters on period j
π_j is the probability of serving a variant on period j

Example

Variant	Cumulative Conv Rate	Inverse probability weighting Conv Rate
Default	9.6%	$\left(\frac{900}{90\%}+\frac{450}{50\%}\right)\div\left(\frac{9000}{90\%}+\frac{5000}{50\%}\right)=\mathbf{9.5\%}$
Alternative	9.3%	$\left(\frac{105}{10\%}+\frac{455}{50\%}\right)\div\left(\frac{1000}{10\%}+\frac{5000}{50\%}\right)=\mathbf{9.8\%}$

ORACLE

Challenge 6: Testing in Multiple metrics

Business need:
Alternative should reasonably improve several KPIs

Consider multiple pairs of hypotheses:
$H_0^j : p_0(M_j) > p_1(M_j)$ vs $H_1^j : p_0(M_j) < p_1(M_j)$

AND – Alternative should outperform Default in **ALL** KPI’s

- Difficult to achieve
- No corrections are needed for p**-values** assuming winner

OR – Alternative should outperform Default in **AT LEASE ONE** of KPI’s

- Simple to achieve
- Bonferroni-type correction is needed

Gatekeeper procedures

- **Example goal:**
  (Alternative > Default in M1) **OR** (Alternative < Default in M1 at most 1% **AND** Alternative > Default in M2)
- Corrections depend on the procedure

ORACLE

Other challenges

Outlier filtering
- Marketing campaigns
- Extreme purchases
- Bots, crawlers, …

Trends detection
- Seasonality
- Long-term effects
- Novelty effect
- Data window

Factors interaction
- Speed up conclusion in MVT
- Reveal usable knowledge

Segments analysis
- Visitors heterogeneity
- “Actionable” insights

Thank you!