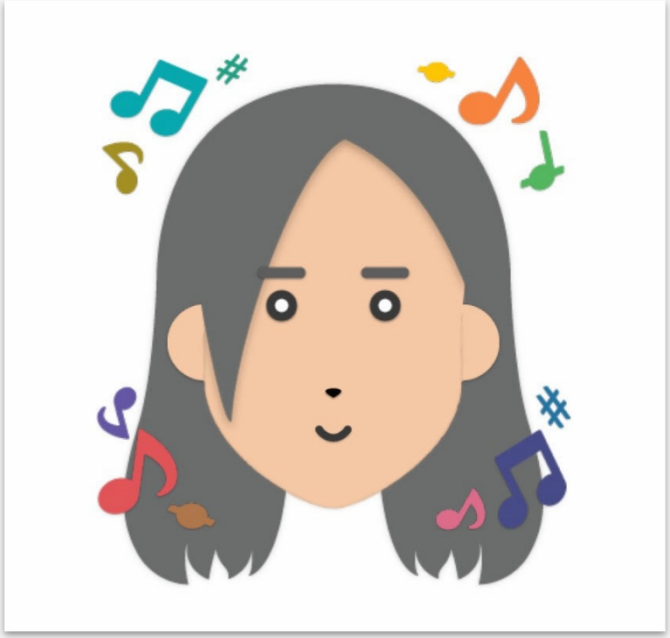
はじめてのA/Bテスト

田中麻衣子
株式会社truestar

TRUESTAR

自己紹介

TRUESTAR



田中麻衣子

- 株式会社truestar（シニアコンサルタント）
- 大学院修士課程卒業後、独立行政法人に勤務
- 2017年より現職
- 現職では、分析業務を中心に担当

Solution Consulting

Operation Consulting	TableauやAlteryxなどのBIツールを駆使し、課題に対し**最適なデータ活用**を実現
Data Analytics	統計解析や機械学習等を用いて、ビジネスの意思決定に役立つ**データ分析**を実施
Spatial Analytics	位置情報データと地図情報や地理統計データを組み合わせた**空間分析ソリューション**を提供
System Solution	分析基盤の構築や自社課題にカスタマイズした**ツールの開発**など、様々な環境に適合可能なソリューションを提供
Strategic Planning	定量的な分析結果に裏付けされた**戦略立案**とその**効果検証**を実施

Platform

Hawkeye Viewer	**マーケットレポート**の統合プラットフォームサービス

本日のお話したいこと

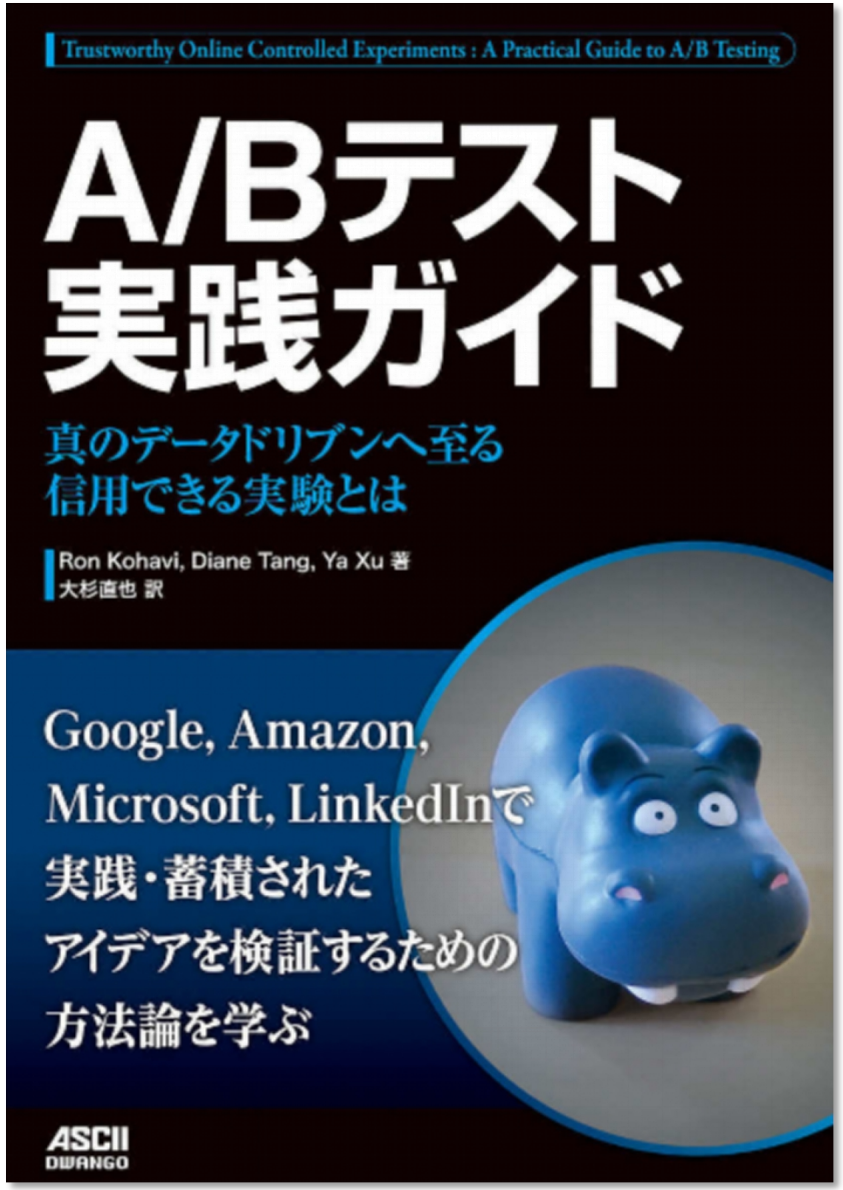
TRUESTAR

- アジェンダ
  1. Ron Kohavi （ロン・コハヴィ）著「A/Bテスト実践ガイド　真のデータドリブンへ至る信用できる実験とは」のご紹介
  2. A/Bテストとは？
  3. **一般的なA/Bテストの流れ（注意点・お役立ちツールの紹介）**
  4. 統計学的検定とその解釈

- お話できないこと
  1. システム実装系のお話
  2. 統計学的検定の詳細・具体的なやり方（Excel操作、R・Pythonコードの書き方等）
  3. 経営者レベルでの意思決定のお話

書籍のご案内

TRUESTAR



2021年3月アスキードワンゴより出版

Ron Kohavi （ロン・コハヴィ）

- Airbnbのヴァイスプレジデント兼技術フェローである
- Microsoftでコーポレートヴァイスプレジデント兼技術フェローをしていた時期に執筆本書執筆
- それ以前はAmazonでデータマイニングとパーソナライゼーションのディレクターであった
- スタンフォード大学でコンピューターサイエンスの博士号を取得
- 彼の論文は40,000件以上引用され、そのうち3つはコンピューターサイエンスで最も引用された論文のトップ1,000に入っている。

本書紹介文：

- 経営層やリーダー、データサイエンティスト、エンジニア、アナリスト、プロダクトマネージャーなど、企業の全関係者を対象として、その概念から実験を実施するためのプロセス、よくある落とし穴、オンライン実験をスケールさせるために必要なことまで、A/Bテスト、コントロール実験の詳細を解説する。
- **Google, Amazon, Microsoft, LinkedInなどで実際に実践された経験から得られた教訓や落とし穴などがふんだんに盛り込まれており**、データ駆動型の文化の確立、科学的な意思決定に必要なスキルを獲得できる。

Amazonより引用

本日は、本書をベースにA/Bテストの流れや陥りがちな罠について一緒に勉強していきましょう！

A/Bテストとは？

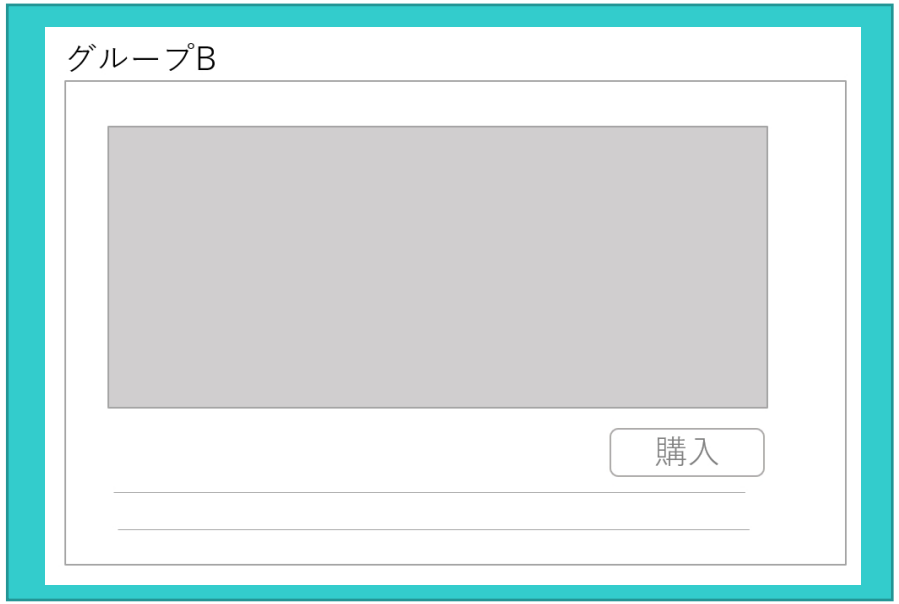
TRUESTAR

概要

- AとBの2つのグループを作成し、各グループのパフォーマンスを測定してどちらがよりパフォーマンスが優れているかをテストすること（＝コントロール実験）

- 実験後、2つのグループ間で差が生じた場合、その差は偶々ではなかったか、バイアスがなかったか等検証したうえで、ビジネスの改善施策へと活用する

テスト場面例：ECサイトで割引クーポン施策を実施するかどうか検討。購入確認画面にクーポンコードの入力欄を追加した場合の影響を評価したい



グループA：介入群

グループB：コントロール群

A/Bテストとは？ – bingの成功事例

TRUESTAR

2012年、ある社員が広告のヘッドライン表示のアイディアを提案

ただ、この変更は優先順位「低」で半年以上バックログに……

ある日、コードが簡単だったため、エンジニアが変更を実施

↓その方法が……

ランダムに選ばれた一部ユーザーに新しタイトルとレイアウトを表示し、そうでないユーザーには古いレイアウトを表示

Bingの収益は12%増

（当時米国内だけで年間換算1億ドル以上の収益増加）

一方でレイアウトの変更は大きな変更ではなく、ユーザー体験を不愉快にするようなものではなかった

本書：Kindle の位置No.387より抜粋

A/Bテストとは？

TRUESTAR

A/Bテストの活用範囲例（オンライン）

- ユーザーインターフェース（UI）
- アルゴリズム（ 検索、 広告、 パーソナライゼーション、 リコメンデーション など）
- レイテンシ（速度）やパフォーマンス
- コンテンツ管理システム
- 顧客サポートシステムの変更など

実験チャンネル例

- Webサイト
- デスク トップ アプリケーション
- モバイル アプリケーション
- 電子メール など
- 実店舗

もっともオーソドックスな効果検証

企業例

amazon Booking.com ebay facebook G in Microsoft

NETFLIX YAHOO! JAPAN Uber Yandex

一般的なA/Bテストの流れ

TRUESTAR

1. 実験仮説の設定
2. グループ分けをするランダム化単位の決定
3. サンプルサイズの決定
4. テスト期間の決定
5. 実験の実施とデータ収集
6. 統計学的検定（or ベイジアンテスト）
7. 結果の解釈と意思決定

一般的なA/Bテストの流れ（本日）

TRUESTAR

1. 実験仮説の設定
2. グループ分けをするランダム化単位の決定
3. サンプルサイズの決定
4. テスト期間の決定
5. 実験の実施とデータ収集
6. 統計学的検定（or ベイジアンテスト）
7. 結果の解釈と意思決定

仮想ECサイトでA/Bテストの流れをおってみます

①A/Bテストでなにを明らかにしたいのか？（仮説設定）

TRUESTAR

明らかにしたいこと

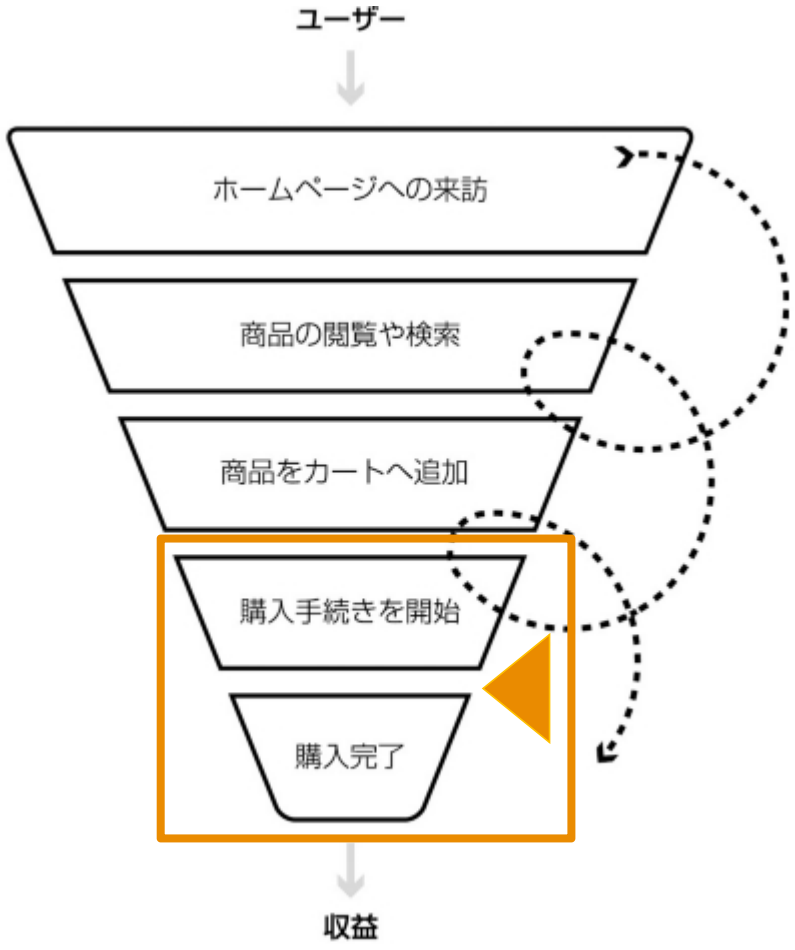
- 購入確認画面にクーポンコードの入力欄を追加した場合の影響を評価したい

仮説設定までの考え方

1. オンラインショッピングのファネルの中で、どこのファネルを改善したいか考える
2. 今回は「購入手続き開始」→「購入完了」の間に問題設定を置く（右図）
3. 具体的な目標指標を設定する
   - I. 総合評価指標（Overall Evaluation Criterion）の決定
     - OECは短期的（実験期間中）に測定可能で、**かつ長期的な戦略目標を推進する原因となるもの（例えば収益）**
   - **II. キーメトリクス**（ゴール指標）の決定
     - OECに紐づく実験期間中に測定可能な指標
   - **➡仮ECサイトの場合、「購入完了率」をキーメトリクスと設定する**
4. 分母を考える
   - I. サイトを訪問したすべてのユーザー：変更が行われた購入確認ページに訪問しなかったユーザーが含まれノイズが大きい
   - II. 購入完了したユーザーのみ：購入を完了したユーザーへの影響ではなく、購入額に影響したかどうかを見たいので適切ではない
   - **III.購入手続きを開始したユーザーのみ：これが一番適切**

導かれた実験仮説

購入確認ページにクーポンコード入力欄を追加すると、購入手続きを開始したユーザーの購入完了率が上昇するのではないか？



図：本書：Kindle の位置No.999より抜粋

②グループ分けをするランダム化単位を決定する

TRUESTAR

ランダム化単位の考え方

- ランダム化単位＝ランダムサンプリングを実施する単位のこと
- 分析単位を同じにすると分析がしやすいので、ランダム化の単位はキーメトリクスに合わせることを推奨する
  - 今回、購入手続きを開始したユーザーの購入完了率が指標であるため、**ユーザーレベルをランダム化単位（サンプリング単位）にする**
  - その他ランダム化単位例：ページレベル、セッションレベル

どの母集団をターゲットにテストするのか？

- 全ユーザー
- 言語特性
- 地理的特徴
- プラットフォーム
- デバイス種類
- 属性（性別・年齢）　など

ターゲットを絞る＝ある特性をもつユーザーのみを対象にすること➡
「誰の何を」明らかにしたいのか？
によってターゲットは異なる

③サンプルサイズってどう決めるの？

TRUESTAR

重用なこと

サンプルサイズは結果の精度に影響を与える

➡サンプルサイズが大きい方がA/B間の小さな差も見逃さず結果に自信が持てる

ただし‥‥

ユーザーがどのように反応するかわからないような大きな変更の場合は、大きいとリスクが高い（ユーザー離脱など）！

つまり‥‥

➡セオリーはあるものの、リスクはビジネス的に判断しなければいけない

どこまで精度を求めるか、どこまで小さな差を検出したいのか

厳密性（コスト）がビジネス的にいくらの価値に換算できるのか

セオリーの話（統計的な話）を少しだけ

③サンプルサイズってどう決めるの？

TRUESTAR

サンプルサイズの決定のポイント

以下のポイントを満たすようにサンプルサイズを決定すると、統計学的に信頼できるテスト結果が得られる

	基準値	キーワード
A/Bの差が偶然生じる割合（または信頼区間）	5%未満（信頼区間95%）	有意水準・統計的有意差
A/Bの差として検出したい最小の差分	ビジネス判断	最小検出可能効果
A/Bの差として検出したい最小の差分の検出力	80%～	統計的検出力

最小検出可能効果

弊社は大企業だから、A/B間の1%の購入完了率の差も見逃せないな…

サンプルサイズ例

		最小検出可能効果		
ベースライン キーメトリクス （平均購入完了率）		5%	10%	25%
	1%	626,231	157,697	25,768
	10%	56,885	14,313	2,332
	25%	18,929	4,754	770

有意水準α＝0.05、統計的検出力(1-β)=0.8

③サンプルサイズってどう決めるの？

TRUESTAR

サンプルサイズ計算の

お役立ちサイトは最後にご紹介

④テスト期間はどう決めるの？

TRUESTAR

重要なこと

- テスト期間はA/Bテストの内容によって変わる

デザインの一部変更	比較的短期でOK
SNSの追加機能や販促促進機能の導入効果	ある程度長期期間で確認する必要がある

- テスト期間の「介入効果」（＝グループAに変更したUIを見せ“介入”している）を精確に検証するためには、長期/短期でそれぞれ留意する点がある

例：「UIの変更」が「売上」に影響があったように見えたが、実はクリスマスシーズン（季節性要素）で放っておいても売上が上がっていた（交絡効果の影響があった）

④テスト期間はどう決めるの？

TRUESTAR

期間	交絡効果	概要
短期テスト	曜日効果	平日と週末でユーザー行動の分布が異なります。テスト期間は最低1週間あると安心
	祝日効果・季節性	祝日や長期休暇タイミングなど、ユーザー行動の分布が異なる
	プライマシー効果とノベルティ効果	実験初期の効果が大きくなったり小さくなったりする傾向のこと UI変更後物珍しさにクリック数が伸びても時間の経過とともにクリック数は減少する場合がある
長期テスト（一般的に3ヶ月以上）	ユーザーの学習効果	ユーザーが変化に対応するとユーザー行動が変化する。 例：実装からある程度時間がたってから有用性に気づく場合や、初期に熱心に探索を行うなど、このような場合ユーザーが最終的な均衡点に到達するまで、長期効果は短期効果とは異なる等
	ネットワーク効果	ネットワークの広がりによって新機能や新アルゴリズムの評価は変わるためどの時点で評価するか検討が必要
	体験と測定の遅延	ユーザー全体が介入効果（新機能の搭載など）を体験するまでには時間差がある場合がある
	エコシステムの変化	エコシステム内の多くのことが時間の経過とともに変化し、ユーザーへの反応に影響（他の新機能のローンチ、競合の状況、政府の方針、ソフトウェアの劣化）

④テスト期間はどう決めるの？

TRUESTAR

変更による**介入効果のみ**を取り出す方法

実験方法・分析方法	概要
コホート分析	• 分析前に、安定したユーザーのコホートを構築しておき、このコホートに対する短期的な効果と長期的な効果を分析する • 構築しておくコホートは潜在的なユーザー全体を代表している必要があるため、ユーザーをサブグループに層別化し（エンゲージメント高中低など）介入効果に重みづけ調整などをする • そうすることで、**コホートによる介入の影響などを取り除いた介入効果が検証できる可能性が高い**
時間をずらした処理（A/Aテスト）	• いわゆるA/A分析のひとつ • **実験と実験の間を空け、A/A分析を実施することで時間による介入を取り除く**
ホールドバック/リバース実験	• 時間がない時 • ホールドバック：90%のユーザーに介入した後、数週間の間10%のユーザーをコントロール群に維持する • リバース：100%のユーザーに介入開始した数週間後に10%のユーザーをコントロール群に戻す • 短期間で時間による介入効果への影響を取り除くが、ユーザーへの影響が大きい
差分の差分法（Difference in Difference）	• 介入前後の差分を介入されたグループとされなかったグループでそれぞれ算出 • さらにグループ間でその差をとる・・・つまり2回差をとるのでDID

実験デザインまとめ

TRUESTAR

1. ランダム化単位はユーザー
2. 全てのユーザーを対象にして、購入確認ページに訪れたユーザーを分析する
3. 1ユーザーあたりの購入完了率の1%以上の変化に80%の検出力を持たせるとして、サンプルサイズを決定
4. サンプルサイズが集まるまでの最小4日実験（これは仮定でサンプルサイズが集まる日数はサイトの流入量などに大きく左右されます。）
5. ただし、曜日効果を考慮するため一週間実施しプライマシー効果などを検出した場合は実験期間を長くする。

⑤実験の実施とデータの収集

TRUESTAR

計測器の実装等は、システムや設備投資の話なので割愛

ただし…集まったデータを確認する中で、サンプル比率が意図通りになっていない場合は注意（サンプルデザインは介入効果とは独立していなければいけない！）

例：

- サンプル比率をグループA：グループB＝1：1に設定したのに、0.99より小さい、または1.01より大きい
- サンプルサイズを決める際に、p値<0.05で設計したにもかかわらず、AB間の差の検定をしたら、p値が0.001以下（あまりにも結果が良すぎる）

⑤実験の実施とデータの収集

TRUESTAR

考えられる可能性例

- ブラウザのリダイレクト
- 介入群とコントロール群でのパフォーマンスの違い
  介入群に遅延が生じ指標に影響を与えている
- ボット
- 特定のブラウザやブラウザバージョンでのバグ
- 計測の欠損
- 残留効果または繰り越し効果（いわゆるバグの混入）
- 時間効果
  メール送付などで、適切にグループはランダム化されていたが、メールの送付順番が介入群、コントロール群など固定されており受け取る時間帯にそもそも時間差があった

日ごとに確認して異常な出来事がなかったか、特定のブラウザの挙動がおかしくなかったか、新規ユーザーと再訪者の比率などこのあたりはやはり基礎集計で異常値チェック

⑥統計学検定をしてみる

TRUESTAR

▼収集されたデータ

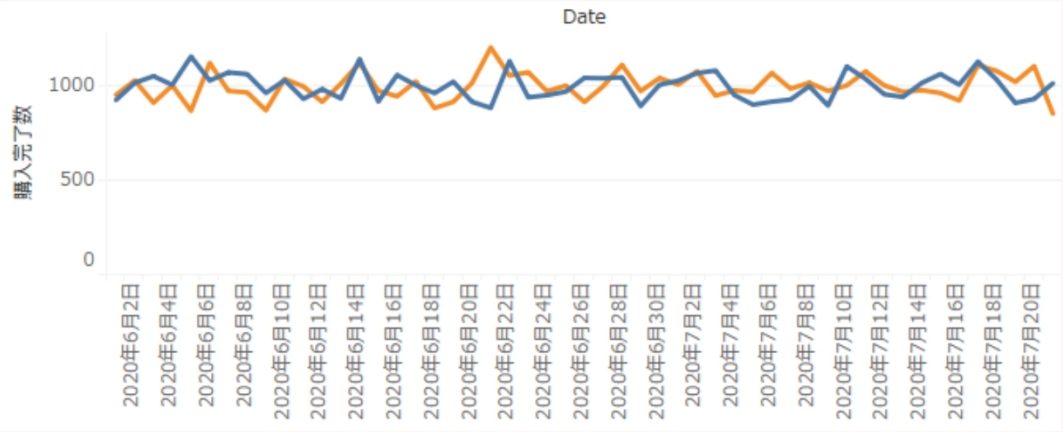
	A	B	C	D	E
1	date	landingPagePath	合計UU	離脱数	購入完了数
2	2020/6/1	ページA	1021	95	926
3	2020/6/1	ページB	1053	100	953
4	2020/6/2	ページA	1127	110	1017
5	2020/6/2	ページB	1129	100	1029
6	2020/6/3	ページA	1145	93	1052
7	2020/6/3	ページB	1008	97	911
8	2020/6/4	ページA	1108	102	1006
9	2020/6/4	ページB	1107	104	1003

▼集計と可視化

データ概要

Landing Page Path	合計uu	購入完了数	平均CVR
ページA	56,117	51,010	90.85%
ページB	55,948	51,001	91.13%

購入完了数推移



AページとBページ、どちらがいいでしょうか？

平均CVRが高いので、ページBでしょうか？

⑥統計学検定をしてみる

TRUESTAR

▼収集されたデータ

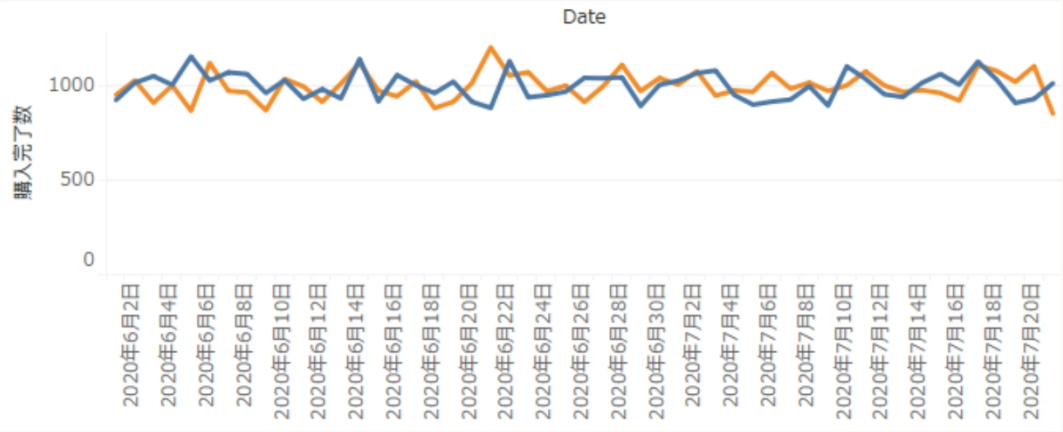
	A	B	C	D	E
1	date	landingPagePath	合計UU	離脱数	購入完了数
2	2020/6/1	ページA	1021	95	926
3	2020/6/1	ページB	1053	100	953
4	2020/6/2	ページA	1127	110	1017
5	2020/6/2	ページB	1129	100	1029
6	2020/6/3	ページA	1145	93	1052
7	2020/6/3	ページB	1008	97	911
8	2020/6/4	ページA	1108	102	1006
9	2020/6/4	ページB	1107	104	1003

▼集計と可視化

データ概要

Landing Page Path	合計uu	購入完了数	平均CVR
ページA	56,117	51,010	90.85%
ページB	55,948	51,001	91.13%

購入完了数推移



「AとBの差が、偶然や誤差とは考えにくく、差に意味がある」というために必要なステップが統計学的検定

⑥統計学検定をしてみる

TRUESTAR

統計学的検定の考え方

1. 仮説の生成（AがBするとCに**なる**）
2. 反証仮説（AがBするとCに**ならない**）
3. 検定

   AがBするとCに**なる**頻度が高すぎる…これは反証仮説を支持できないぞ…
4. 反証仮説が棄却される
5. 仮説が支持される

※事象が生じる頻度が100回中96回以上だと、偶々とは言えないという基準がある（$p<0.05$）

⑥統計学検定をしてみる

TRUESTAR

今回のケース

- 仮説：

  購入確認ページにクーポンコード入力欄を追加すると、購入手続きを開始したユーザーの購入完了率が上昇する

- 反証仮説：

  購入確認ページにクーポンコード入力欄を追加しても、購入手続きを開始したユーザーの購入完了率は上昇しない

- 基準：

  p値<0.05（ABグループ間で生じた差がたまたまである割合は5%以下）

TRUESTAR

統計学的検定の種類

	2つのグループの違い	2つ以上のグループの違い
元の分布が正規分布を仮定する	t検定・カイ二乗検定	ANOVA検定
元の分布が正規分布を仮定しない	ウィルコクソン検定	クラスカル・ウォリス検定

変数種類

連続変数 → 平均の差 → t検定

離散変数 → **割合の差** → カイ二乗検定

TRUESTAR

検定の実施

- 目的変数：購入を完了しているか否か（TRUE/FALSE）
- 説明変数：LandingPagepath（Aページor Bページ）

※カイ二乗検定はExcelでも*Evan's Awesome A/B Tools*でもできます　*（最後のページに操作方法説明します）*

サマリ　確率分布　ピボットテーブル　比率　カイ二乗値への寄与度 (%)　差分

カイ二乗値	P値	自由度	効果量 (Cohen's w)	検出力	タイプ2エラーの確率	行の数
2.2921884755	0.13002646975	1	0.0045226206838	0.32806715068	0.67193284932	112,065

※今回Exploratoryというツールを利用
※また、カイ二乗検定はデータの持ち方を加工しなければいけません
※加工方法の考え方はhttps://exploratory.io/note/GMq1Qom5tS/A-B-wHL0xqZ0tmのブログで確認できます

TRUESTAR

検定の結果を解釈する

サマリ　確率分布　ピボットテーブル　比率　カイ二乗値への寄与度 (%)　差分

カイ二乗値	P値	自由度	効果量 (Cohen's w)	検出力	タイプ2エラーの確率	行の数
2.2921884755	0.13002646975	1	0.0045226206838	0.32806715068	0.67193284932	112,065

- p=0.13（目標はp<0.05）つまり偶然ABグループに差が生じる割合が13%
- 反証仮説が棄却できない＝**ABグループの差に意味があるとは言えない**

すみません、
結果は有意ではありませんでした

それはどういう意味？？
次何すればいいの？？

⑥統計学検定をしてみる

TRUESTAR

- ただし、有意でないからといって意味がないわけではなく、差の検出力が低い（サンプルサイズが小さい）場合もある→ちなみに今回はこのケース
- その上で、よく指摘される一般的な統計学的検定のデメリット

1	テスト開始前に、サンプルサイズの検討などある**程度統計的知識**が必要な作業を要す
2	サンプルが集まってからでないと検定ができないので、**タイムラグ**が生じる
3	検定結果は直観的に**理解しづらい**（p値とは…から説明しなければいけない）
4	**有意か有意ではないかの白黒の判断**しかできないので、ビジネス的にネクストステップが決めづらい

ビジネス的なメリットが大きいという理由で
最近は**ベイジアンA/Bテスト**というベイズ推定を用いたテストを採用するケースも

⑥統計学検定をしてみる―結論

TRUESTAR

しかし…検定の結果を施策の実用性と統計学的有意性の掛け算で解釈するべし

“覚えておくべき 重要なことは、結果から明確な答えが得 られなくても、決断しなければならないことがあり得ることである。そのような状況では、どのような要因を考慮しているのか、特にそれらが実用的で統計的に有意な境界線にどのように変換されるのかを明確にしておく必要がある。” (Kindle の位置No.1252)

検定結果	実用性	解釈
有意	実用性あり	ローンチの決定に動き出せる
有意	実用性低	ローンチに対するコストパフォーマンスを検討する必要がある
有意でない	実用性あり	テストを再度実施すべきで、結果の精度を上げるためにより大きな検出力テストを再度実施すべき
有意でない	実用性低	施策にあまり意味はない

今回の例はクーポンを表示しても、しなくても購入完了数に差があるとは言えず、更にクーポンコードの管理や割引率の設定、運用など、その他コストがかさむため、実施は見送るべき…という一旦結論になりそう

⑦意思決定をする

TRUESTAR

A/Bの結果を踏まえ、本当に既存UIやシステムに手を加えるのか？

- 改修によってトレードオフになる指標がないか？

  局所的改善をしても大局的にマイナスになるケースはないか？

- 改修コストが回収できるのか？
- ローンチ後の保守運用コストは改修に見合うのか？

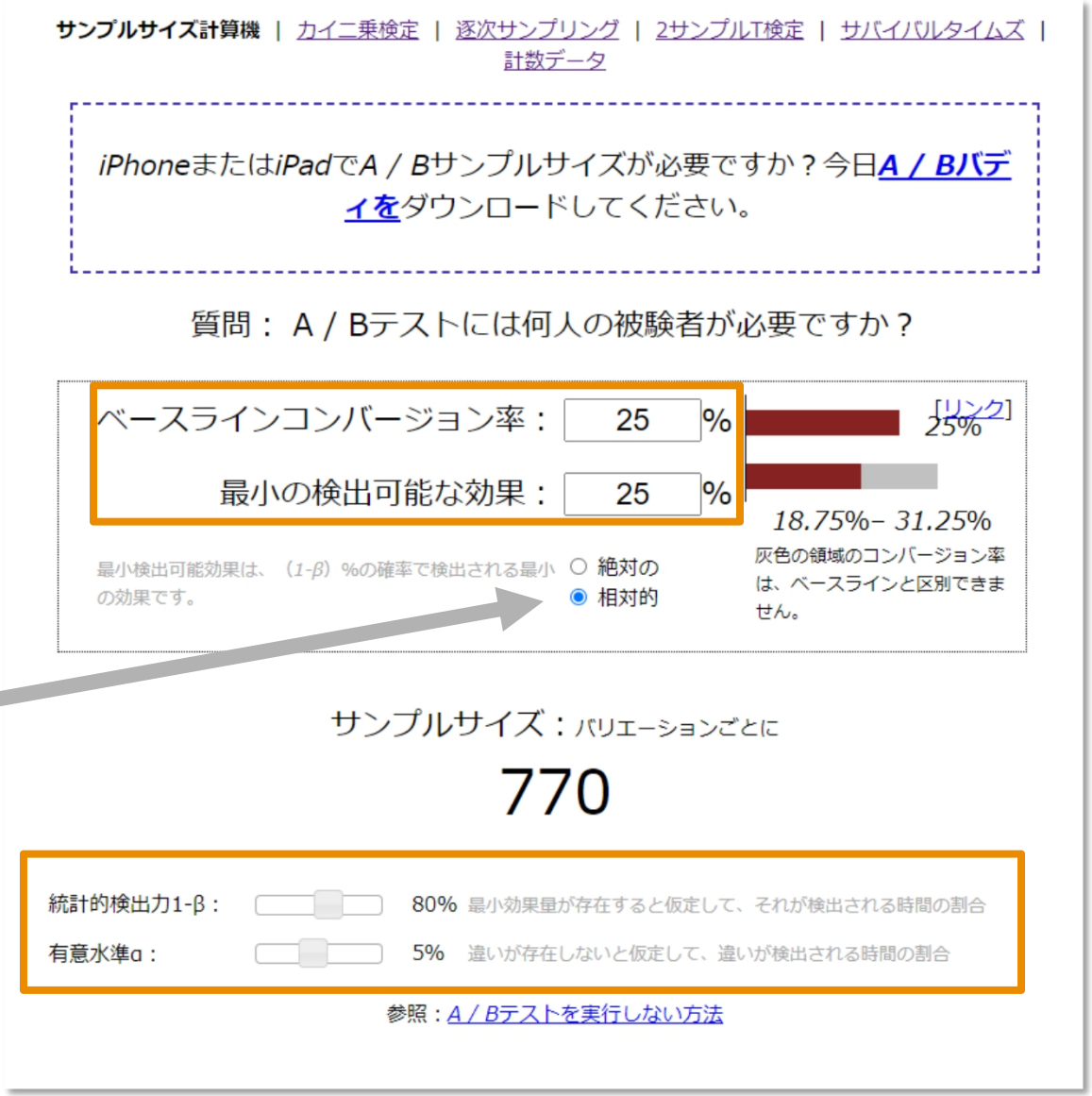
サンプルサイズ便利計算サイト

TRUESTAR

■検定でカイ二乗検定をする場合

Evan's Awesome A/B Tools

https://www.evanmiller.org/ab-testing/sample-size.html



- コンバージョン率に基づく、「相対的な変化」測定するように設定（Exコンバージョン率の1%の変化）

■検定をt検定で行う場合

https://keisan.casio.jp/exec/user/1490704651

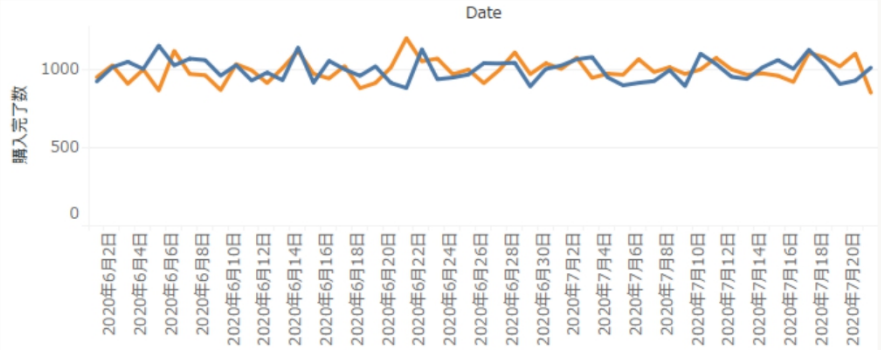
補足：Evan's Awesome A/B Toolsでカイ二乗検定をする方法

データ概要

Landing Page Path	合計uu	購入完了数	平均 CR
ページA	56,117	51,010	90.85%
ページB	55,948	51,001	91.13%

Landing Page Path
ページA
ページB

購入完了数推移



Evan's Awesome A/B Tools (home):

Sample Size Calculator | **Chi-Squared Test** | Sequential Sampling | 2 Sample T-Test | Survival Times | Count Data

Question: Does the rate of success differ across two groups?

	# successes	# trials	Confidence interval [clear] [link]
Sample 1:	51010	56117	*90.7% – 91.1%*
Sample 2:	51001	55948	*90.9% – 91.4%*

Verdict:

No significant difference

(p = 0.13)

Confidence level: 95%

If the experiment is repeated many times, the confidence level is the percent of the time each sample's success rate will fall within the reported confidence interval.

It is also the percent of the time no difference will be detected between the two groups, assuming no difference exists.

- 統計学検定が有意の場合でも、p値と併せて「効果量」を報告するのがベスト
- 効果量＝検定で有意な差を見つけた時にその差の大きさを表す指標（０～１の値）
- 目安：0.1＝小、0.3＝中、0.5＝大

効果量 (Cohen's w)
0.0045226206838

ブログのご紹介

- A/Bテストに関してブログを書いています
  - はじめてのA/Bテスト前篇

    https://blog.truestar.co.jp/exploratory/20210524/44719/

  - はじめてのA/Bテスト後編

    https://blog.truestar.co.jp/exploratory/20210612/45074/

https//www.truestar.co.jp/

Ebisu Business Tower 4F | 1-19-19 Ebisu | Shibuya-ku | Tokyo 150-0013 | Japan
Tel: 03 5422 6561 | Fax: 03 5422 6562 | e-mail: info@truestar.co.jp