A/B Testing Analysis Using Frequentist and Bayesian Inference

---

1. Executive Overview

This project evaluates whether a newly designed landing page improves user conversion rates relative to the existing version. A randomized controlled A/B test was simulated, involving approximately 294,000 users evenly split between treatment (new page) and control (old page) groups over a 23-day period.

The **primary metric** is the user conversion rate, defined as the proportion of visitors completing a predefined action (e.g., account sign-up). The experiment is designed with a **5% significance level** and **80% statistical power** to detect a **minimum detectable effect (MDE)** of **0.75 percentage points**.

This report applies both **Frequentist** and **Bayesian** inferential frameworks. The Frequentist approach relies on classical hypothesis testing using proportions, confidence intervals, and p-values. The Bayesian approach models uncertainty through posterior distributions and evaluates the probability that the treatment outperforms the control under the observed data.

The goal is to derive evidence-based recommendations that guide landing page deployment, optimizing for both **statistical rigor** and **business decision-making** under uncertainty.

---

2. Introduction: Concepts, Context, and Roadmap

This project investigates a central question in conversion rate optimization:
Does a redesigned landing page improve user conversion compared to the existing version?

To explore this, we simulate an experiment using a publicly available dataset from Udacity's A/B testing course. Although the experiment was not conducted firsthand, we reconstruct its context and structure analytically to reflect what a proper real-world experiment would look like.

Report Structure:

This report is structured as follows:

- Theoretical Background
  A concise overview of A/B testing principles, hypothesis testing frameworks, and key statistical concepts.
- Simulated Experiment Reconstruction
  Details of the simulated setup, including randomization, assumptions, success criteria, and sample size planning.
- Data Preparation
  Initial exploration, cleaning, and transformation of the dataset to ensure consistency, remove anomalies, and construct relevant variables for analysis.
- Frequentist Analysis
  Application of classical hypothesis tests, confidence intervals, and validation of assumptions.

- Bayesian Analysis
  Probabilistic modelling using Beta priors and posterior inference to quantify uncertainty and support decision-making.
- Expanded Analysis: Regional Performance
  A deeper, country-specific, analysis conducted to uncover any regional performance variations
- Comparative Observations
  Examination of agreements and divergences between Frequentist and Bayesian results, with practical implications.
- Conclusion and Recommendations
  Synthesis of findings and guidance for future testing and deployment

---

3. Theoretical Background

A/B testing is a controlled experimental methodology used to compare two variants—typically a control and a treatment group—to determine whether a proposed change has a statistically significant effect on a key performance metric. In digital contexts, this typically involves randomly assigning users to different experiences and measuring conversion outcomes to infer causal impact.

Key Concepts:

- Randomization
  Random assignment of users to groups eliminates selection bias and ensures group comparability under the assumption of exchangeability. This enables valid causal inference.

- Hypothesis Testing
  - The **null hypothesis (H_0)** assumes no difference in conversion rates between control and treatment.

  - The **alternative hypothesis (H_1)** posits that a difference exists, either in a one-sided or two-sided form depending on the business question.

- Type I and Type II Errors
  - **A Type I error (α)** occurs when a true null hypothesis is incorrectly rejected **(false positive).** It is typically controlled at **5%**.

  - **A Type II error (β)** occurs when a false null hypothesis fails to be rejected **(false negative).** The **statistical power** of a test**,** defined as **1−β,** is often targeted at **80%.**

- Minimum Detectable Effect (MDE)
  The smallest effect size considered practically significant. It is used in conjunction with power and significance level to determine the required sample size for the experiment.

- Frequentist Inference
  - Relies on **test statistics** (e.g., z-test for proportions), **p-values**, and **confidence intervals** to assess evidence against **H_0**.

- A **p-value** quantifies the probability of observing a result as extreme as the one obtained, under the assumption that $\mathbf{H_o}$ is true.

- A **confidence interval** provides a range of plausible values for the true effect size**,** with a specified confidence level (e.g., 95%) under repeated sampling.

- Bayesian Inference
  - Models uncertainty using **probability distributions** over unknown parameters, updating **prior** beliefs with observed data via **Bayes' theorem**.

  - The **posterior distribution** reflects updated beliefs about the true conversion rates and the treatment effect.

  - **Credible intervals** represent ranges within which the parameter lies with a specified posterior probability (e.g., 95%).

  - Decision-making can be guided by evaluating the **posterior probability** that the treatment outperforms control, yielding more intuitive probabilistic conclusions.

Primary Metric:

- Conversion Rate
  The proportion of users in each group who complete a predefined target action (e.g., account sign-up). This is a binary outcome summarized as a proportion, suitable for both proportion-based hypothesis testing and Beta-Binomial modelling in the Bayesian framework.

---

4. Simulated Experiment Reconstruction

Goal of the Experiment:

To evaluate whether a **new landing page** increases the **conversion rate** compared to the existing **old landing page**.

A. Experiment Overview

Feature	**Value**
Population Size	294,478 total user visits
Groups	Treatment (new page): 147,276 users Control (old page): 147,202 users
Conversion Metric	Binary converted column (1 if converted, else 0)
Conversion Rate	Control: 12.04% Treatment: 11.89%
Date Range	Jan 2, 2017 → Jan 24, 2017 (23 days)

Feature	Value
Traffic Allocation	~50% to each group
Unit of Analysis	Individual user visits (user-level randomization)
Landing Pages	old_page, new_page

B. Experiment Setup

Component	Details
Objective	Test if the new landing page leads to significantly different conversion than the old one.
Primary Metric	Conversion rate (binary outcome: converted)
Secondary Metrics	Session duration (if available), bounce rate (not in dataset, could be suggested for future experiments)
Hypotheses	**- Null**: $p_{new} = p_{old}$ **- Alternative**: $p_{new} \neq p_{old}$ (two-sided) **- Significance Level:** $\alpha = 0.05$ **- Power:** 80% **- Minimum Detectable Effect (MDE):** 0.75 percentage points **- Randomization Unit:** user_id **- Randomization Mechanism:** 50-50 random assignment to treatment or control group **- Assignment Logic:** Users assigned to either control (old page) or treatment (new page) group, and saw corresponding landing page

C. Sample Size Calculation Assumptions

- **Baseline Conversion Rate (control)**: ~12.04%
- **Target Lift**: +0.75 percentage points (e.g., increase from 12.04% to 12.79%)
- **Significance Level**: 0.05
- **Power**: 0.8

Note: Sample size was not planned prospectively in this dataset. These values reflect what would have been required had the experiment been prospectively designed under the listed assumptions.

D. Experiment Validity Checks

These are steps we **simulate having conducted** before analysing the dataset:

- **Sanity checks**:
    - Group sizes are nearly equal (147k each)
    - No major group/page mismatches (although ~3,800 mismatches exist)
- **Equal variance assumption**: Assume not grossly violated
- **Randomization check**: We assume that randomization was effective, with balanced distributions across observable covariates (not directly validated in this dataset)
- **No significant peeking or early stopping**: Assume one-time evaluation

Note: Approximately 3,800 users experienced group/page mismatches. These will be excluded or analysed separately.

E. Realistic Simulated Experiment Narrative

"We simulate a controlled online **A/B test** conducted over 23 days (Jan 2–24, 2017) to evaluate a new landing page design. Users were randomly assigned to either the **treatment group (new page)** or the **control group (old page)**, and their conversions were recorded as **binary** outcomes. The experiment targeted a **minimum detectable effect** of **0.75 percentage points**, with **80% power** and a **5% significance level**. In total, **294,478 user sessions** were collected and form the basis of our analysis."

---

5. Data Preparation

Exploratory Data Analysis (EDA):

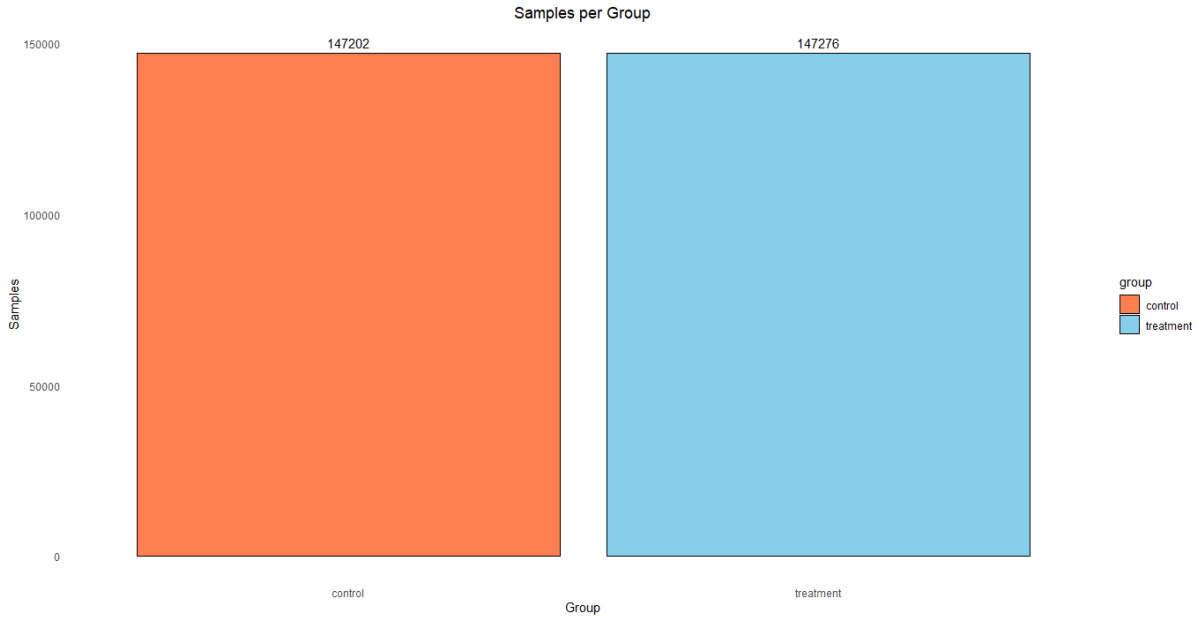
- Data Structure
    - **user_id:** Unique identifier per user.
    - **timestamp:** Date and time of the user's visit.
    - **group:** Assigned experimental condition (control or treatment).
    - **landing_page:** Page version seen (old or new).
    - **converted:** Binary indicator of conversion (1 = converted, 0 = not converted).
- Data summary
    - The A/B test spanned 23 days (January 2–24, 2017).
    - The control and treatment groups each contain ~145,000 users (147,202 and 147,276, respectively).

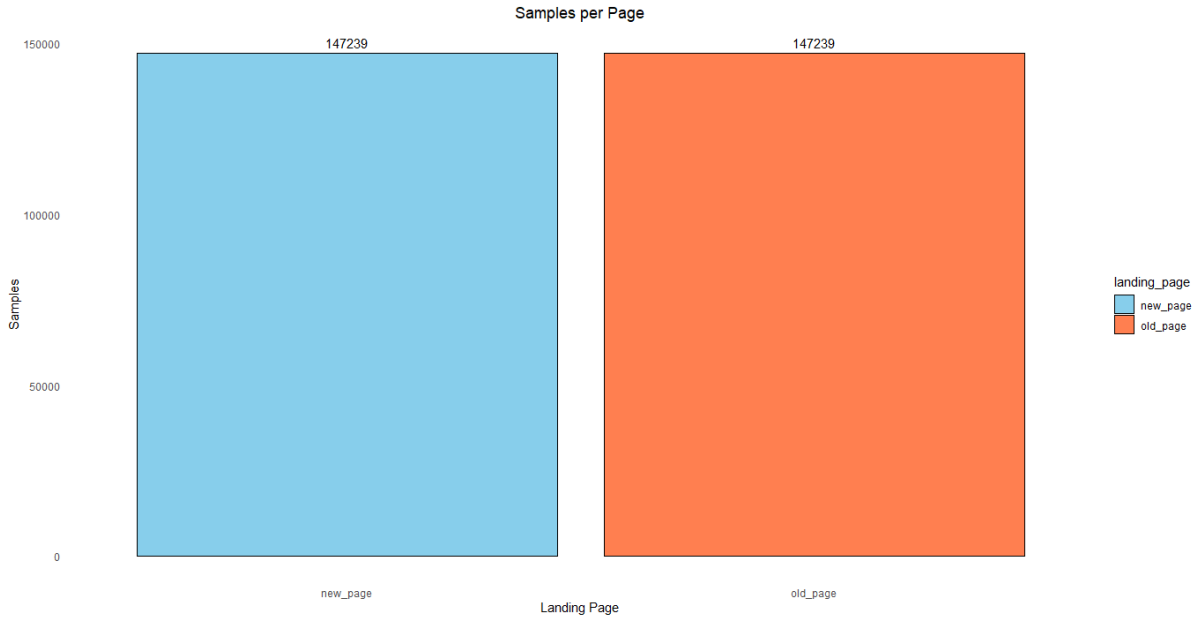
- Both the old and new landing pages appear 147,239 times, indicating group-page mismatches.

- The overall conversion rate across all users is approximately 11.96%.

- Missing Values
  No missing values were detected during preliminary checks.

- Initial Group and Landing Page Distributions
  Visual inspection of the group and landing page distributions revealed imbalances due to mismatches between assigned groups and observed landing pages.





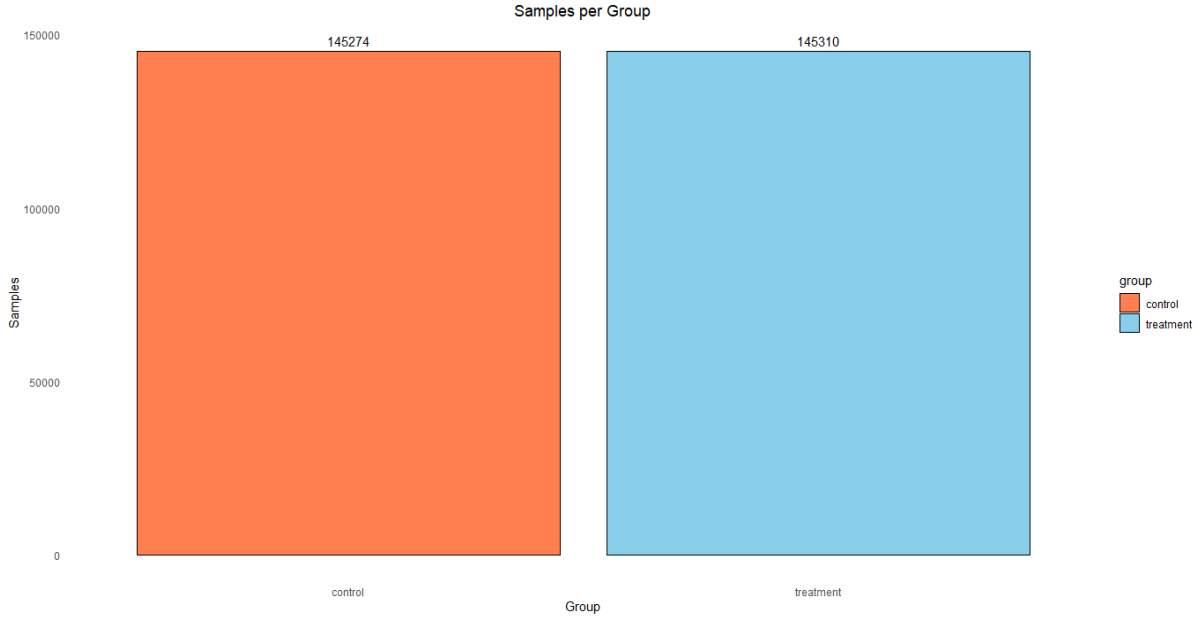
Data Cleaning & Feature Engineering:

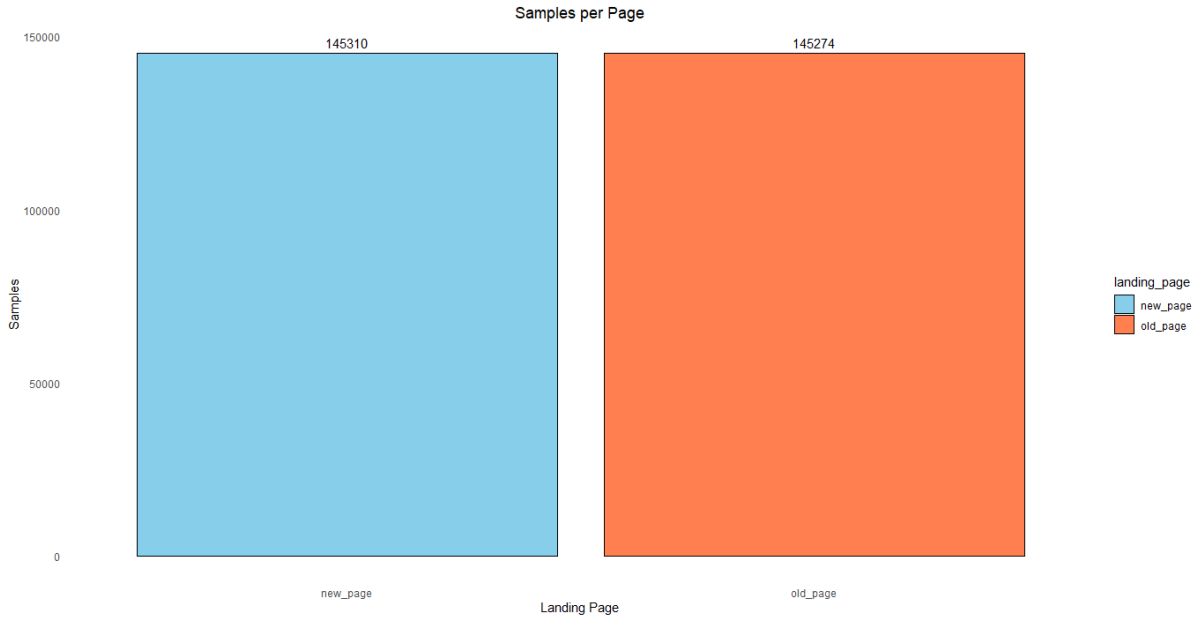
- Mismatch Removal
  1,928 control-group users landed on the new page, and 1,965 treatment-group users landed on the old page. These mismatches were excluded to preserve the integrity of the experimental design.

- Duplicate Removal
  A duplicate entry was found for user_id 773192. One instance was removed to preserve user-level uniqueness.

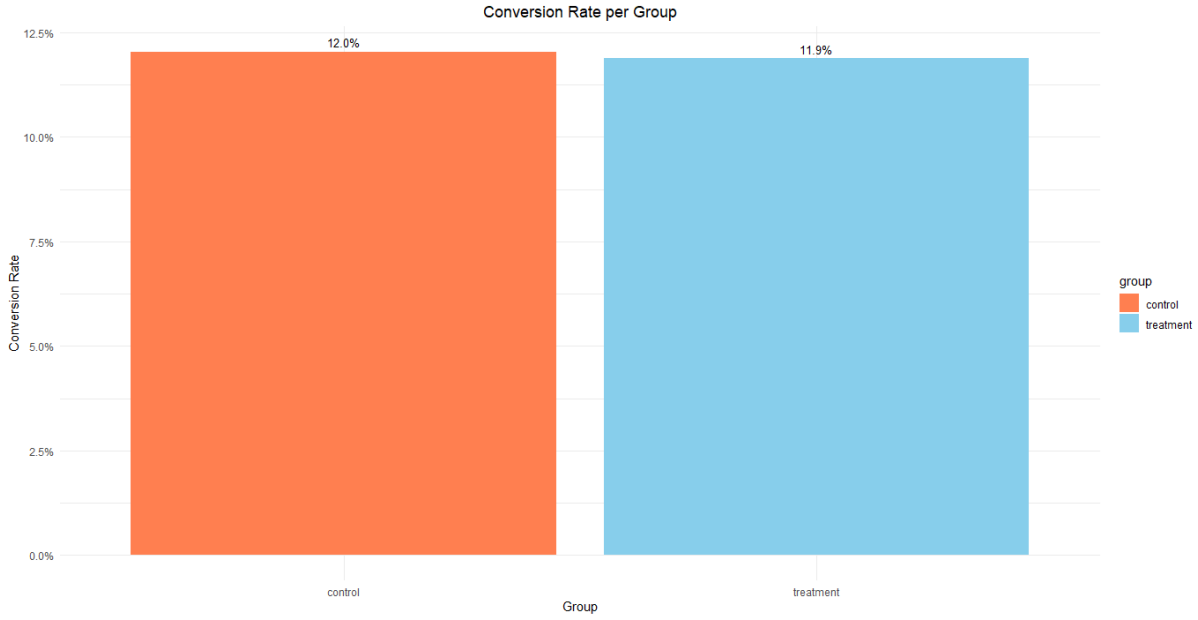
- Variance Stability Across Groups
  The outcome variance is consistent between the treatment and control groups. This suggests that the assumption of equal variance, required for methods like the pooled z-test, is not meaningfully violated. It also provides indirect evidence that randomization was effective, as it did not result in substantially different user types being assigned to each group.

- Post-cleaning Distributions
  User allocation between groups remained stable and balanced after mismatch removal. Landing page distributions also reflected correct assignment post-cleaning.





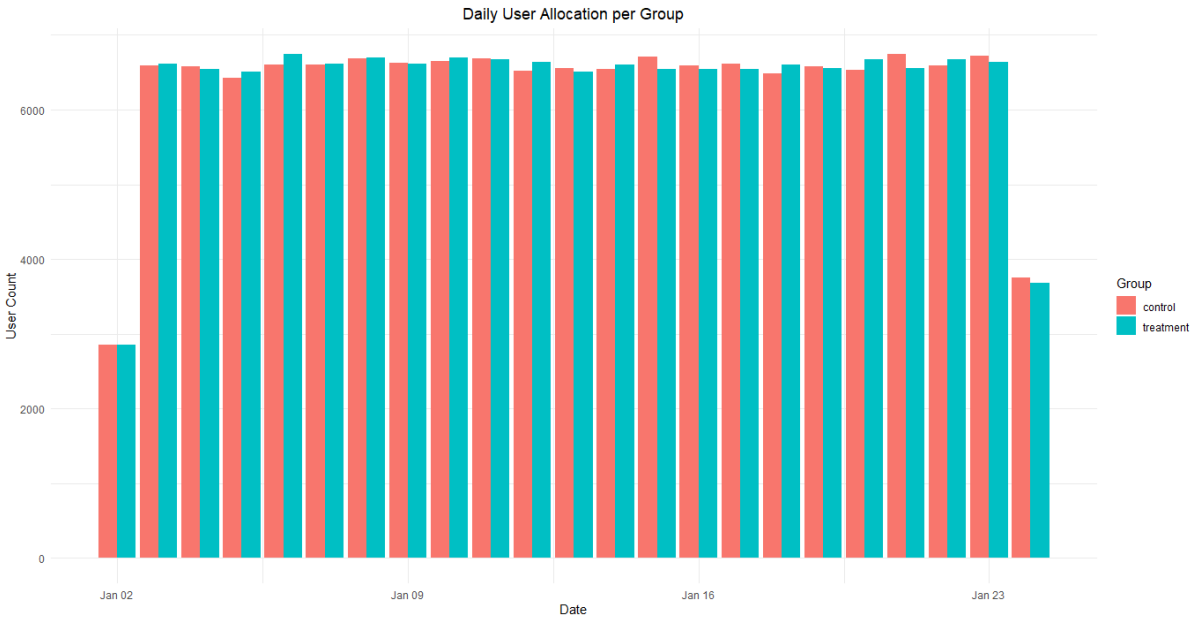
- Conversion Rate Analysis
  The control group had a conversion rate of 12.04%, while the treatment group had a rate of 11.89%. Conversion rates by landing page were identical, confirming consistency after cleaning.

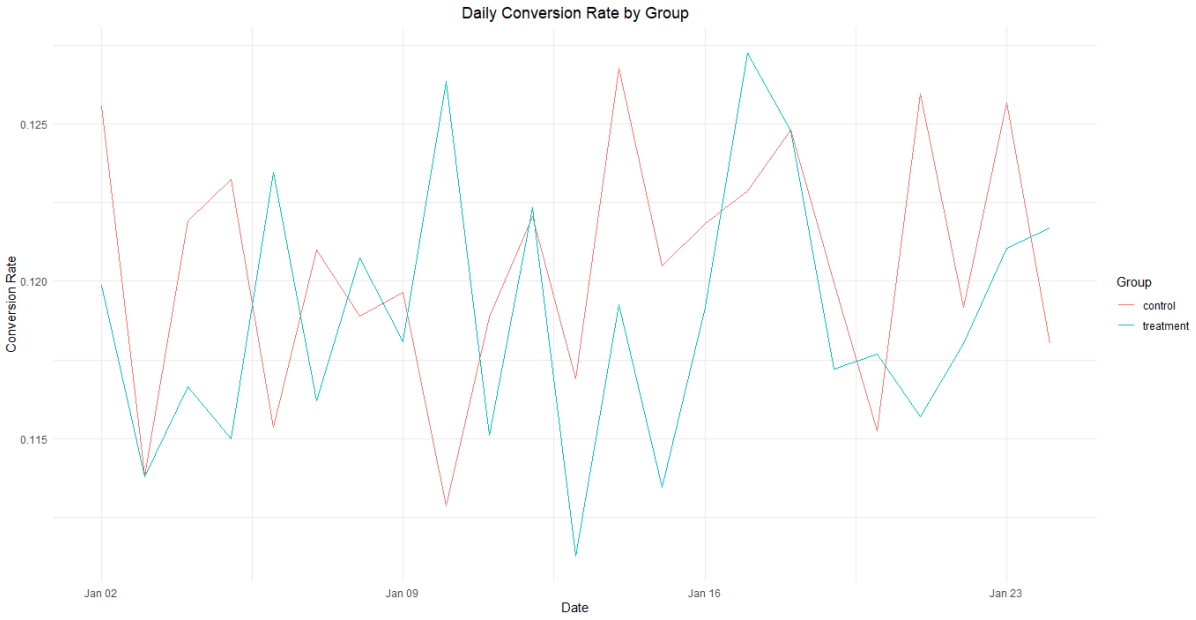


- **Dataset Limitations**
  Due to dataset constraints, no additional user-level covariates (e.g., demographics, device type) are available to assess the randomization balance beyond group sizes and conversion variance. This limits the ability to fully rule out potential confounding, though data integrity checks (duplicate removal, timestamp consistency) support the validity of the analysis.

- Temporal Trends
  Daily group allocation remained consistent across the experiment. Conversion rates fluctuated over time: the control group peaked on January 14 and bottomed on January 10.

The treatment group peaked on January 17 and was lowest on January 13. These fluctuations appear random and do not indicate systematic bias.





---

6. Frequentist Analysis

Define the Problem and Formulate Hypotheses:

- Research Question
  Test if the new landing page leads to significantly different conversion than the old one.

- Stating Hypotheses

- **Null hypothesis (H_0):** The new landing page leads to no different conversion than the old one.
- **Alternative hypothesis (H_1):** The new landing page leads to significantly different conversion than the old one.

Power Analysis Setup:

- **Control conversion rate (p_1):** 12.04% (0.1204)
- **Treatment conversion rate (p_2):** 11.89% (0.1189)
- **Sample sizes:** ~145,274 (control), ~145,311 (treatment)
- **Minimum Detectable Effect (MDE):** 0.75% absolute uplift planned (from 12% to 12.75%)
- **Significance level (α):** 0.05 (two-sided)
- **Power:** 80%
- **Required sample size per group:** ≈ 30,226
- **Actual sample size per group**: ≈ 145,000+ (well above minimum)

Observed results:

- **Observed difference**: treatment conversion is 0.15% lower than control.
- **Pooled proportion:** pp ≈ 0.1196
- **Standard error:** SE ≈ 0.0012
- **Test statistic:** z ≈ −1.31
- **Corresponding two-sided p-value:** ≈ 0.19
- **Confidence Intervals:** [−0.00393, +0.00078]

Interpretation:

- Despite large sample sizes and sufficient statistical power to detect the pre-specified minimum detectable effect (MDE), the observed uplift in conversion was **negative** (−0.15%) and **not statistically significant**.
- The **95% confidence interval** for the difference in conversion rates **(−0.39% to +0.09%)** includes **zero**, indicating that the observed effect may be due to chance.
- The **p-value exceeds the 5% significance threshold**, providing **no statistical evidence** to reject the null hypothesis of equal conversion rates between control and treatment groups.

- **The treatment does not lead to a measurable improvement in conversion**. If anything, it may have a slight detrimental effect — but this effect is both **small in magnitude** and **statistically indistinguishable from random variation**.

---

7. Bayesian Analysis

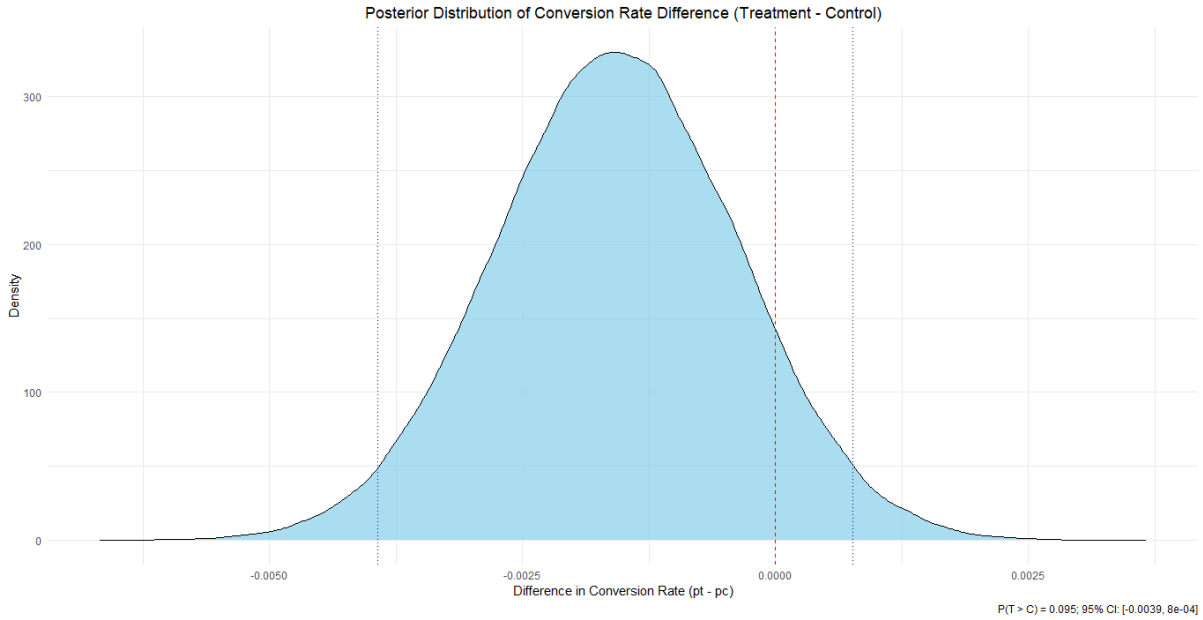
- Set prior distributions
  We assume uninformative priors for both group conversion rates:
  - $pc \sim Beta(1,1)$
  - $pt \sim Beta(1,1)$

Note: We use Beta(1,1) as a flat, uninformative prior, reflecting no prior knowledge and letting the data dominate. In smaller samples, a weakly informative prior can improve stability and guard against implausible estimates. Here, large sample sizes justify a noncommittal choice.

- Posterior distributions
  Upon observing the data:
  - $pc|data \sim Beta(17490, 127786)$
  - $pt|data \sim Beta(17265, 128047)$

- Sample from posteriors
  To approximate the posterior distribution of the difference in conversion rates, we draw:
  - 100000 samples from Beta (17490, 127786) for the control group.
  - 100000 samples from Beta (17265, 128047) for the treatment group.

- **Compute the elementwise difference:** This provides a Monte Carlo approximation of the posterior distribution of the **treatment effect**, $\delta = pt - pc$. The posterior distribution of the difference in conversion rates is shown below.



- **Summary statistics**
    - Posterior probability that treatment outperforms control: P(pt > pc) ≈ 9.5%
    - 95% credible interval for pt - pc: [-0.00391, +0.00076]
- **Region of Practical Equivalence (ROPE):**
  We define a **ROPE** around zero to represent effect sizes considered practically negligible
    - **ROPE:** [−0.003, +0.003]
    - Posterior mass within **ROPE**: ~82%

Note: The ROPE threshold was selected based on practical significance—i.e., changes smaller than ±0.3 percentage points are not actionable given cost, risk, and impact. This allows us to assess not just *whether* there's a difference, but whether it's **meaningful**.

Interpretation:

- Only ~9.5% of the posterior distribution lies above zero, providing **weak evidence** in favour of the treatment.
- The **95% credible interval [−0.00391, +0.00076**] is **centred slightly below zero** and includes both practically negative and practically null effects. This implies that the treatment may be slightly worse than control or have no real effect.
- Given the high posterior mass of ~82% within this ROPE, we have strong evidence that the new page's effect is too small to be meaningful, providing a quantitative basis for the decision not to deploy it.

---

8. Expanded Analysis: Regional Performance

While the overall A/B test showed no significant difference, a deeper, country-specific analysis was conducted to uncover any regional performance variations. Due to sample size constraints in the Canadian segment (14,499 samples), a standard Frequentist analysis would have been underpowered. To provide a consistent and meaningful analysis across all regions, a Bayesian posterior comparison was performed instead.

Our analysis of the new page's conversion rate in each country revealed the following:

- US: The probability that the new page is better than the old page is 6.5%. The 95% credible interval is [-0.00499, 0.00062], and the ROPE coverage is 71.9%.
- UK: The probability that the new page is better than the old page is 68.3%. The 95% credible interval is [-0.00358, 0.00588], and the ROPE coverage is 73.4%.
- CA: The probability that the new page is better than the old page is 9.6%. The 95% credible interval is [-0.01733, 0.00341], and the ROPE coverage is 20.1%.

These results provide a nuanced perspective that goes beyond a simple go/no-go decision. While the UK shows a higher probability of the new page being better, the credible intervals for all countries still overlap with zero, and the ROPE coverage for the US and UK is high, suggesting any potential lift is likely too small to be practically significant at a global scale. The low ROPE coverage in Canada is a result of the smaller sample size, which leads to a wider credible interval and higher uncertainty.

---

9. Comparative Observations

Post-Hoc Observations and Power Limitations (Frequentist Analysis):

Frequentist hypothesis testing revealed no statistically significant difference between the treatment and control groups. The observed −0.15% difference in conversion rates failed to meet the pre-specified minimum detectable effect (MDE) of 0.75%, and the associated p-value of 0.19 exceeds the conventional 5% significance threshold.

The 95% confidence interval for the difference in conversion rates [−0.00393, +0.00078] includes zero and skews negative, reinforcing the lack of statistical evidence for a meaningful treatment effect. Despite ample sample size and power to detect small effects, this result suggests that the observed variation may be attributable to random chance rather than a systematic benefit of the new landing page.

From a decision-making perspective, these results provide no justification to implement the treatment. The evidence supports a conservative stance: the new page is unlikely to outperform the current version, and it may even slightly underperform. In light of these findings, the prudent course is to halt rollout and re-evaluate the proposed changes.

Post-Hoc Observations and Decision Limits (Bayesian Analysis):

Posterior inference yielded a 9.5% probability that treatment outperforms control, and a 95% credible interval of [−0.00391, +0.00076], which includes zero and skews negative.

Using a ROPE of ±0.3%, approximately 82% of the posterior mass falls within this region. This suggests that the treatment effect is likely inconsequential from a business perspective.

In Bayesian terms, this does not provide evidence in favour of the treatment. Rather, it provides moderate evidence for practical equivalence — i.e., that the new page is no better (and possibly worse) than the old.

Although Bayesian methods avoid binary hypothesis testing, this result supports the same operational decision: there is insufficient evidence to justify deploying the new page. The minor −0.15% observed drop, combined with high ROPE coverage and low probability of improvement, points toward conservatism and re-evaluation rather than rollout.

---

10. Conclusion and Recommendations

Decision:

Do not deploy the new landing page at this time.

Reasoning:

- Both **Frequentist** (p-value = 0.190, CI includes 0) and **Bayesian** (Pr(pB > pA) ≈ 9.5%) results suggest that the new page **does not improve conversion** — and may even underperform.
- A large share of the **posterior distribution falls within the ROPE** (Region of Practical Equivalence), indicating a high probability that the difference is practically negligible.
- The observed lift (~–0.15%) is **below the Minimum Detectable Effect (0.75%)**, rendering the experiment **underpowered to detect such a small shift**.
- The expanded regional analysis supports this conclusion but also provides a more strategic path forward. The data indicates that the new page showed the most promising results in the **UK**, where there is a **68.3% probability** that its conversion rate is higher than the control. However, given the high ROPE coverage, this difference may not be practically meaningful.

Next Steps:

- Redesign and retest: Consider changes to layout, content, or CTA strategy, and run a new A/B test if the new design justifies it.
- Broaden your metric scope: Track secondary metrics (e.g., session duration, click-throughs, bounce rate) to better understand user engagement beyond binary conversion.
- Cost-benefit assessment: Evaluate whether pursuing further design iterations is worth the **engineering cost, opportunity cost, and potential lift**.
- A **targeted follow-up test** in the UK market is recommended with a larger sample size to more confidently measure the potential uplift. For the US and Canada, it is recommended to keep the original landing page and explore alternative designs to better address regional user preferences.