**HOW PROMOTION 1 PROVED TO BE THE MOST EFFECTIVE FOR INCREASING AVERAGE WEEKLY SALES**  
FAST FOOD MARKETING CAMPAIGN A\B TEST RESULTS

# | 1. Scenario

A fast-food chain plans to add a new item to its menu. However, they are still undecided between three possible marketing campaigns for promoting the new product. In order to determine which promotion has the greatest effect on sales, the new item is introduced at locations in several randomly selected markets. A different promotion is used at each location, and the weekly sales of the new item are recorded for the first four weeks.

## 1.1. Goal

Evaluate A/B testing results and decide which marketing strategy works the best. Read more about this under the section Target Metric.

## 1.2. Columns

These are the columns included in the original table.  
*Note: the dataset is aggregated by LocationID, PromotionID and week.*

| ***market\_id*** | Unique identifier for market |
| --- | --- |
| ***market\_size*** | Size of market area by sales |
| ***location\_id*** | Unique identifier for store location |
| ***age\_of\_store*** | Age of store in years |
| ***promotion*** | One of three promotions that were tested |
| ***week*** | One of four weeks when the promotions were run |
| ***sales\_in\_thousands*** | Sales amount for a specific *LocationID*, *Promotion*, and *week* |

## 1.3. Target metric

One of the key aspects of this A/B test is identifying the most relevant metrics that allow to compare the effectiveness of the three promotions.

But how? The most crucial metric in this case is sales, as it directly reflects the impact of each marketing campaign. Specifically, the variable sales\_in\_thousands provides a measure of the week-by-week total revenue generated for each store location under a given promotion.

To evaluate the overall performance trends of each promotion, independent of the location, the weekly sales data was aggregated and averaged. This allows for a more stable and comprehensive comparison between the promotions, reducing the impact of short-term fluctuations and providing a clearer picture of which strategy delivers the best long-term results. From this point forward, this metric will be referred to as average weekly sales.

# | 2. **Calculations**

The Two-Sample T-Test from the [Evan Miller A/B test calculator](https://www.evanmiller.org/ab-testing/) was selected to perform the statistical comparison between the promotions.  
This test is designed to compare the means of two samples, and determine whether there is a significant difference between them considering a pre-determined confidence interval.  
Since there are three marketing campaigns, three pairwise comparisons will be conducted to identify which one has the highestaverage weekly sales.

For each comparison, the Null Hypothesis *(H0)* states that the two sample means are equal, indicating no significant difference, while the Alternative Hypothesis *(Ha)* suggests that a significant difference exists between them.  
To mitigate the risk of a multiple testing problem, which increases the chances of getting a Type I error, a 99% confidence level (α) is used instead of the standard 95%.   
A Type I error occurs when the Null Hypothesis (H₀) is rejected, even though it is actually true. In simpler terms, it means concluding there is a significant effect or difference when, in reality, there isn’t one. It is also known as false positive.

A SQL query was written and executed in BigQuery\* to aggregate the necessary data and compute key statistics for the T-test (sample size, mean and standard deviation). The results were then imported into the calculator for the actual comparison.

\*Please check section 4.1 Appendix for more info.

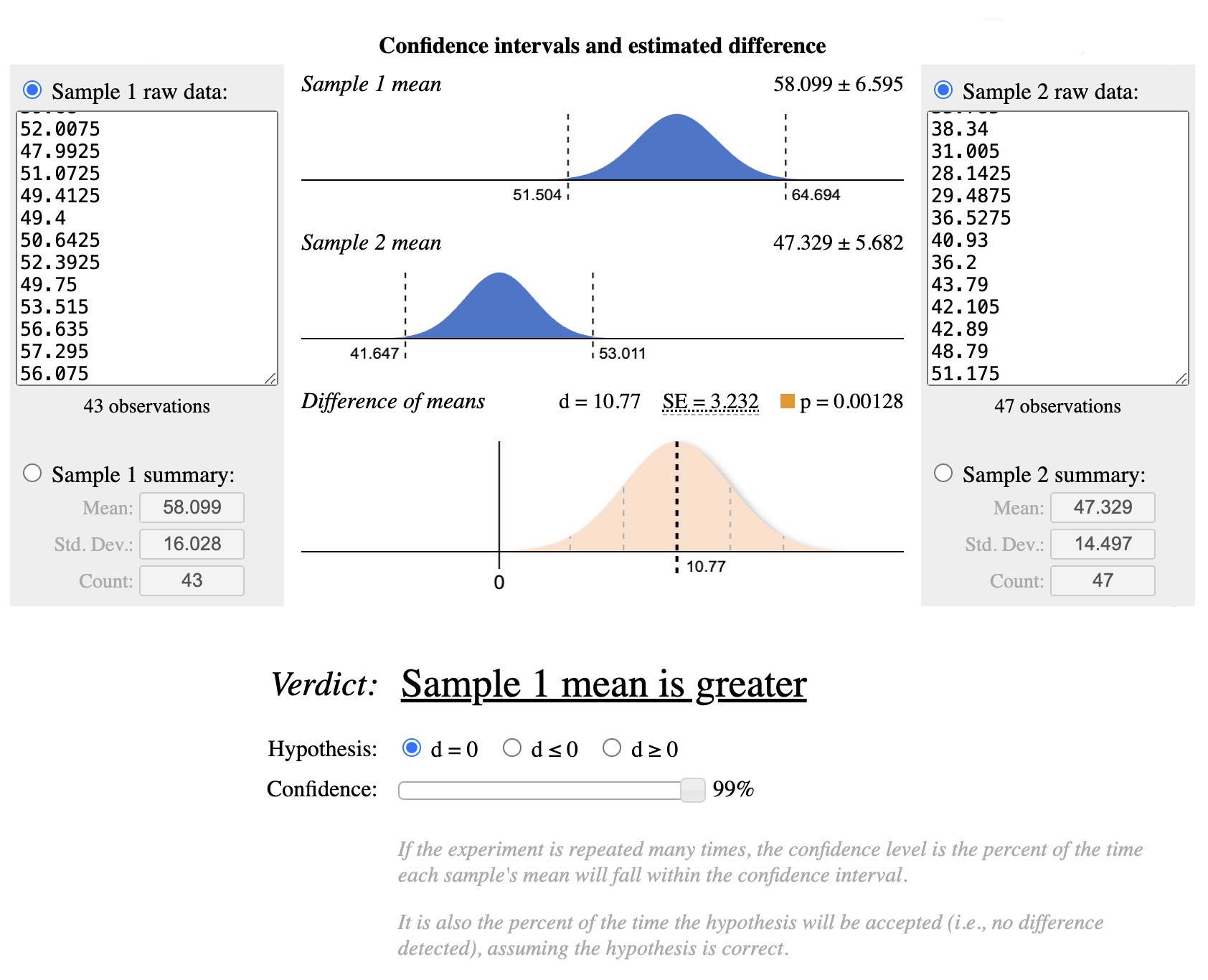
## 2.1. Query results

| **promotion** | **sample\_size** | **mean** | **standard\_deviation** |
| --- | --- | --- | --- |
| 1 | 43 | 58.099 | 16.0282 |
| 2 | 47 | 47.3294 | 14.4971 |
| 3 | 47 | 14.4971 | 16.3839 |

## **2.**2. **p\_1** *(Sample 1, left)* **vs p\_2** *(Sample 2, right)* **comparison**

*H₀:**p\_1**=**p\_2**or p\_1**-**p\_2**= 0*

*Ha:**p\_1**p\_2**or p\_1**-**p\_2 0*



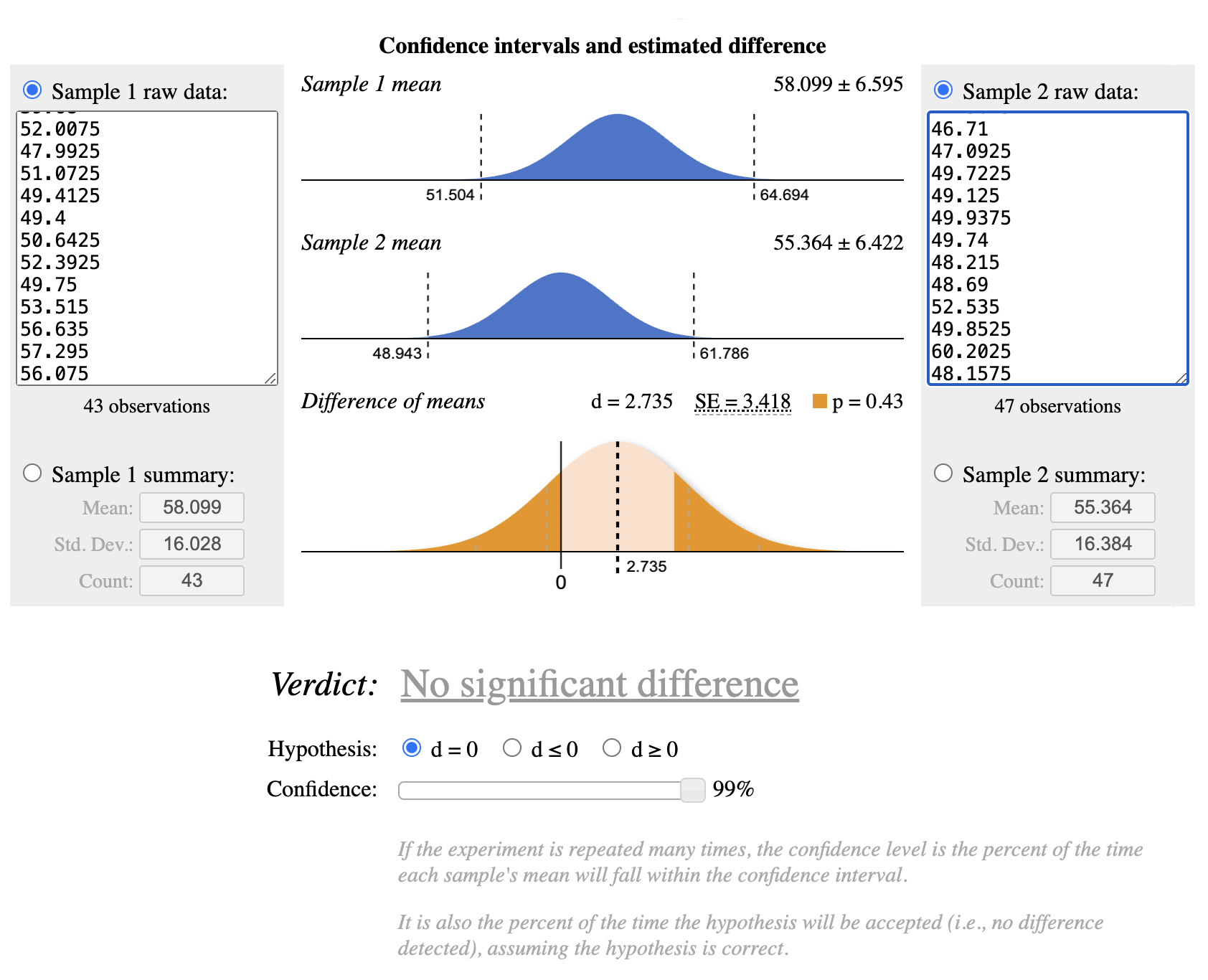
### 2.1.1. Result

With a 99% confidence level (α = 0.01) and a p-value of approximately 0.001, the results indicate a statistically significant difference between the means of p\_1 and p\_2. This low p-value strongly suggests that the observed difference is unlikely to be due to random chance alone.

## **2.2**. **p\_1** *(Sample 1, left)* **vs p\_3** *(Sample 2, right)* **comparison**

*H₀:**p\_1**=**p\_3**or p\_1**-**p\_3**= 0*

*Ha:**p\_1**p\_3**or p\_1**-**p\_3 0*

****

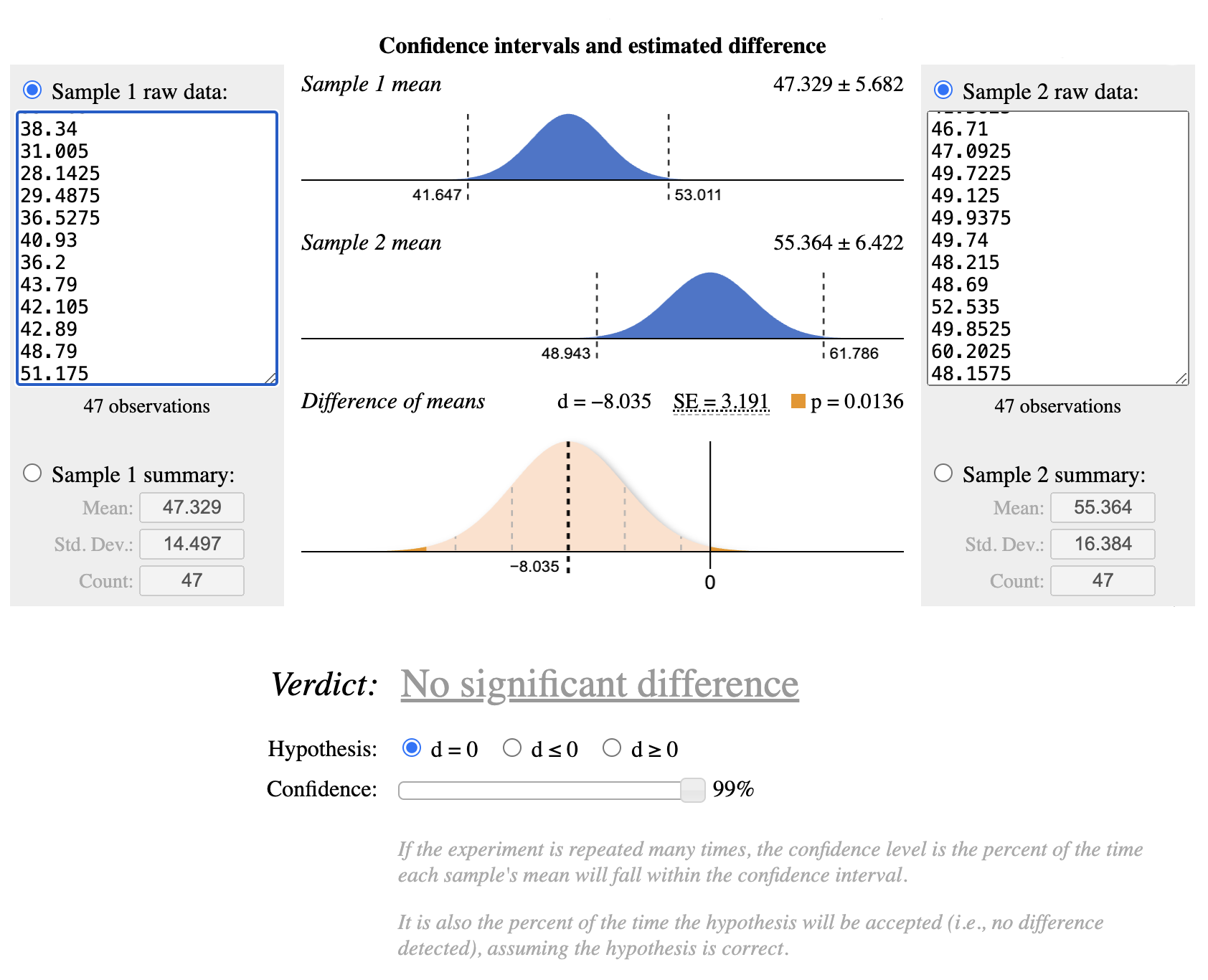
### 2.2.1. Result

With a 99% confidence level (α = 0.01) and a p-value of 0.43, the results indicate no statistically significant difference between the means of p\_1 and p\_3. The confidence intervals of both samples overlap, and the difference in means (d = 2.735) is relatively small compared to the standard error (SE = 3.418). This high p-value suggests that any observed difference is likely due to random variation rather than a true effect.

## **2.3. p\_2** *(Sample 1, left)* **vs p\_3** *(Sample 2, right)* **comparison**

*H₀:**p\_2**=**p\_3**or p\_2**-**p\_3**= 0*

*Ha:**p\_2**p\_3**or p\_2**-**p\_3 0*

****

### 2.3.1. Result

Given that our confidence level is 99% (α = 0.01) and the p-value is 0.0136, the difference between the two groups is not statistically significant at this confidence level.

While the observed difference suggests that p\_3 outperforms p\_2, the evidence is not strong enough to confidently reject the null hypothesis. Therefore, we cannot conclude with high certainty that there is a meaningful difference between these two promotions.

# | 3. **Decision**

From these results, promotion 1 emerges as the most effective in driving higher average weekly sales, particularly when compared to promotion 2.

On the other hand, since promotions 1 and 3 do not show a statistically significant difference, either could be considered a viable option depending on other business factors, such as cost efficiency, marketing strategy, or target audience response.

**Given this information, the recommended course of action is:**

**a.** Prioritize promotion 1, as it has shown a clear advantage over promotion 2.

**b.**  If additional budget is available, consider testing promotion 3 further, as it performed similarly to promotion 1.

**c.**  Avoid using promotion 2 as it performed the worst in terms of sales.

Further analysis with a larger dataset or different confidence levels could help refine these insights and confirm long-term effectiveness.

# | 4. **Appendix**

## 4.1. Query for table

WITH aggregated\_data AS

(

SELECT location\_id,

promotion,

ROUND(AVG(sales\_in\_thousands), 4) AS average\_weekly\_sales

FROM `wa\_marketing\_campaign`

GROUP BY location\_id,

promotion

)

SELECT promotion,

COUNT(location\_id) AS sample\_size,

ROUND(AVG(average\_weekly\_sales), 4) AS mean,

ROUND(STDDEV(average\_weekly\_sales), 4) AS standard\_deviation

FROM aggregated\_data

GROUP BY promotion

This query analyzes the effectiveness of different promotions by comparing the number of locations involved, the average weekly sales, and the variability in sales.

**Step 1:** The CTE *aggregated\_data* calculates the average weekly sales *(AVG(sales\_in\_thousands))* for each location *(location\_id)* and type of promotion *(promotion)*.

**Step 2:** The main query takes the aggregated data and computes the main statistics:

1. *COUNT(location\_id) AS sample\_size* → Number of locations that participated in the promotion.
2. *AVG(average\_weekly\_sales) AS mean* → Average weekly sales across all locations for that promotion.
3. *STDDEV(average\_weekly\_sales) AS standard\_deviatio*n → Standard deviation of sales, to measure the variability of results.

### 

## 4.2. Considerations on α and its impact on the business

As previously mentioned, to mitigate the multiple testing problem, we set a stricter significance level (α = 0.01) instead of the conventional 0.05. This decision significantly reduces the probability of making a Type I error (false positive), ensuring that any detected difference between promotions is genuinely meaningful rather than a result of random variation. However, this stricter threshold also increases the risk of a Type II error (false negative), meaning we might fail to detect a real effect when one actually exists.

### 4.2.1. Business impact of Type I vs. Type II errors

a. Type I Error (False Positive) – If we incorrectly conclude that a promotion is significantly better when it actually isn’t, we risk allocating budget and resources to an ineffective marketing strategy. This could lead to wasted spending and missed opportunities to invest in a truly high-performing promotion.

b. Type II Error (False Negative) – If we fail to detect a real difference between two promotions, we might overlook a genuinely superior strategy, leading to suboptimal decision-making. In this case, we might continue using a less effective promotion, potentially missing out on increased revenue and customer engagement.

### 4.2.2. What’s better?

Given our business context, a Type I error is generally more costly because launching an ineffective promotion at scale could lead to financial losses and misallocated resources. For this reason, we prioritized minimizing false positives by setting a stricter alpha level. However, as seen in the comparison between Promotion 2 and Promotion 3, a slightly less stringent alpha (e.g., 0.05) might have revealed more actionable insights, particularly in borderline cases.

Thus, in future tests, we might consider a context-dependent approach, where the choice of alpha is adjusted based on the potential risks and costs associated with both error types.