Links

https://public.tableau.com/views/GloboxDashboardnew2/GloboxDashboard?:language=en-GB&publish=yes&:display_count=n&:origin=viz_share_link

https://www.statsig.com/calculator?mde=10&bcr=3.9231&twoSided=true&splitRatio=0.5&alpha=0.05&power=0.8

https://www.loom.com/share/cb55800c1a804f798dee27957bc385a1?sid=ceab4489-42b1-46a8-8476-3070778873c0

Globox Database Analysis
Novelty_Effect

SQL Extraction Code

I used a SQL query to evaluate the impact of a new banner on user engagement and spending. Our query gathers data from three tables and measures two key metrics: 'Converted' and 'Total_Spent.' This query lays the groundwork for a thorough A/B test analysis.

-- This SQL query is designed to aggregate user data for A/B testing analysis.

-- The goal is to join the 'users', 'groups', and 'activity' tables

```
-- to obtain a comprehensive dataset that includes user demographics,

-- their group assignment (test/control), and their activity data (amount spent).

SELECT

  -- Select relevant columns from the 'users' table

  u1.id as users_id, -- User ID

  u1.country as users  _country, -- Country of the User

  u1.gender as users_gender, -- Gender of the User

  -- Select relevant columns from the 'groups' table

  g1.device as users_device, -- Device used by the User

  g1.group as test_group, -- Indicates if the user is in the test or control group

  -- Calculate the 'Converted' column:

  -- If the total amount spent is greater than 0, mark it as 'Yes', otherwise 'No'

  CASE

    WHEN SUM(a1.spent) > 0 THEN 'Yes'
```

```
    ELSE 'No'
  END AS Converted,
  -- Calculate 'total_spent' as the sum of the 'spent' column from 'activity'
  -- If there is no data (NULL), replace it with 0
  COALESCE(SUM(a1.spent), 0) AS total_spent
-- Joining Tables
FROM users u1 -- Start with the 'users' table
-- Left join with 'groups' on user ID
LEFT JOIN groups g1
ON u1.id = g1.uid
-- Left join with 'activity' on user ID
LEFT JOIN activity a1
ON u1.id = a1.uid
-- Group the data by user ID, country, gender, device type, and group type
GROUP BY u1.id,
        u1.country,
```

u1.gender,

g1.device,

G1.group

-- Sort the results by user ID in ascending order

ORDER BY u1.id **ASC**;

Statistical Analysis In Spreadsheet

To count the number of users in Group A in a spreadsheet, use the formula **=COUNTIF(E2:E48944, "A")**. This formula scans through cells E2 to E48944 and tallies up the instances where the cell value is "A". Similarly, to count the number of users in Group B, use the formula **=COUNTIF(E2:E48944, "B").**

For counting the number of conversions in Group A and Group B, use the formulas **=COUNTIFS(E2:E48944, "A", G2:G48944,">0") and =COUNTIFS(E2:E48944, "B", G2:G48944,">0")**, respectively. These formulas tally instances where the cell value in column E corresponds to the group and the corresponding value in column G is greater than 0.

To calculate the conversion rate for Group A and Group B, use the formulas **=(COUNTIFS(E2:E48944, "A", G2:G48944,">**

0"))/COUNTIF(E2:E48944, "A") and **=(COUNTIFS(E2:E48944, "B", G2:G48944,"> 0"))/COUNTIF(E2:E48944, "B"**), respectively. These formulas divide the number of conversions (where values in column G are greater than 0) by the total number of users in the respective group (specified in column E).

To calculate the overall conversion rate for all groups combined, use the formula **=COUNTIF(G2:G48944,">0")/COUNT(A2:A48944).** This formula divides the total number of conversions (where values in column G are greater than 0) by the total number of users (non-empty cells in column A).

To calculate the Standard Error (SE) for the difference in conversion rates between two groups in a spreadsheet, use the formula **=SQRT(L10*(1-L10)*(1/L4+1/L5)).** L10 represents the pooled conversion rate, while L4 and L5 represent the total number of users in Group A and Group B respectively.

The formula **=(L8−L9)/L11** calculates the Z-score in a spreadsheet. It takes the difference between the conversion rates of Groups A and B (stored in cells L8 and L9) and divides it by the Standard Error (stored in cell L11). This Z-score is used to determine the statistical significance of the difference between the two groups.

During hypothesis testing, the significance level is crucial and is represented by the value **0.05,** also known as α. This value plays a significant role in determining whether the difference between the conversion rates of Groups A and B is statistically significant. If the

p-value is less than **α=0.05**, which is the threshold, then the null hypothesis is rejected, indicating that the difference is significant. It is imperative to always keep this in mind to ensure accurate results.

To calculate the critical Z-value for a two-tailed test in a spreadsheet, use the formula **NORMSINV(1-(L13/2)),** where **L13** represents the significance level (commonly referred to as α, for example, 0.05) stored in the cell. The **NORMSINV** function returns the Z-value for a given cumulative probability. In a two-tailed test with a significance level of α, the critical Z-value corresponds to a probability of **1-(2α/2).** This critical Z-value is the threshold used to either reject or fail to reject the null hypothesis.

To determine the p-value for a two-tailed Z-test in a spreadsheet, utilize the following formula: **=2*(1-NORMSDIST(ABS(L12))).** The function **'NORMSDIST'** provides the cumulative distribution function (CDF) for the standard normal distribution up to a specific Z-value, which is the absolute value of the Z-score stored in cell **L12**.

This formula evaluates how significant the observed **Z-score** is compared to a standard normal distribution. Multiplying by 2 accounts for the two-tailed nature of the test. The resulting p-value is then compared to a pre-determined significance level (α), typically set at **0.05**, to determine whether or not to reject the null hypothesis. If the p-value is less than α, the null hypothesis is generally rejected,

indicating that the difference between the groups is statistically significant.

The formula **=IF(L15<L13," we reject the null hypothesis", "we fail to reject the null hypothesis")** is used as a decision rule in a spreadsheet for hypothesis testing. It compares the p-value stored in cell **L15** with the significance level α stored in cell **L13.** If the p-value **(L15)** is less than the significance level **(L13)**, the formula returns "we reject the null hypothesis", indicating that the difference between the groups is statistically significant. On the other hand, if the p-value **(L15)** is greater than or equal to the significance level **(L13)**, the formula returns "we fail to reject the null hypothesis". This suggests that there is not enough evidence to conclude that the groups are statistically different. This automated decision-making formula simplifies the interpretation of A/B test results by directly providing the conclusion based on the calculated p-value and significance level.

Use **=L23-L22** to calculate the difference between two values in a spreadsheet, such as conversion rates for A/B testing. This formula represents Δp, the difference between **pB** and **pA**, which are stored in cells L23 and L22, respectively.

To calculate the **Margin of Error (MOE)** in A/B testing and Confidence Intervals using a spreadsheet, you can use the formula: **=(L25*L26).** This formula multiplies the values stored in cells L25 and L26. The MOE is typically calculated by multiplying the Standard Error (which may be stored in cell L25) by the Z-value for the desired confidence level (which may be stored in cell L26). In statistical terms, the formula for MOE is **MOE = Z x SE**. Here, Z represents the Z-value corresponding to the desired confidence level (such as 1.960 for a 95% confidence level), and SE represents the Standard Error.

In a spreadsheet, the formula **=L24-L26** calculates the difference between the values stored in cells L24 and L26. This formula can be used in A/B testing and **Confidence Intervals (CI**) to determine the lower bound of the confidence interval around the difference in conversion rates between Groups B and A. For instance, if L24 contains the point estimate of the difference in conversion rates (Δp), and L26 contains the Margin of Error (MOE), then the formula would compute:

Lower Bound of CI = Δp - MOE

To calculate the sum of values in cells **(L24)** and **(L26)** on a spreadsheet, use the formula **=L24+L26.** This formula is particularly useful in A/B testing and Confidence Intervals (CI), where it can represent the upper

limit of the confidence interval around the difference in conversion rates between Groups A and B.

Upper Bound of CI = Δp + MOE

Use **=AVERAGEIF(E2:E48944, "A", G2:G48944)** to find the average value of column G entries for Group A in a spreadsheet. This helps determine metrics for A/B testing

Use **=AVERAGEIF(E2:E48944, "B", G2:G48944)** to find the average value of column G entries for Group B in a spreadsheet. This helps determine metrics for A/B testing.

To find the standard deviation of Group A values in column G, use the formula `**=(STDEV.S(ARRAYFORMULA(IF(E2:E48944="A", G2:G48944))))**`. This generates an array of only column G values, where column E is "A", then calculates the standard deviation.

To find the standard deviation of Group A values in column G, use the formula `**=(STDEV.S(ARRAYFORMULA(IF(E2:E48944="B", G2:G48944))))**`. This generates an array of only column G values, where column E is "B", then calculates the standard deviation.

To find the **standard error** for the difference in means of two groups with unequal variances in a spreadsheet, use the formula **=SQRT((L38^2/L34)+(L39^2/L35))**. The variables **L38, L39, L34,** and **L35** represent sample standard deviations and sizes. The standard error is then used to calculate the T-statistic for determining statistical significance.

The formula **= ((L38^2/L34 + L39^2/L35)^2) / (((L38^2/L34)^2/(L34-1)) + ((L39^2/L35)^2/(L35-1)))** calculates the **degrees of freedom** for Welch's T-Test in a spreadsheet, which is used when the variances of the two groups are not assumed to be equal.

To find the critical T-value for a two-tailed T-test in a spreadsheet, use the formula **=T.INV.2T(L43, L41),** where L43 is the significance level (usually 0.05) and L41 is the degree of freedom. This value is important for rejecting or failing to reject the null hypothesis in a two-tailed T-test. It serves as a threshold for the T-statistic to reject the null hypothesis at the given significance level and degrees of freedom.

To calculate the two-tailed p-value for a T-test in a spreadsheet, use the formula **=T.DIST.2T(ABS(L42), L41).** The p-value determines whether to reject the null hypothesis. If it's less than the significance level **(alpha = 0.05)**, the difference between the groups is statistically significant.

In a **T-test hypothesis testing** spreadsheet, there is a formula that uses the IF function. It compares the pre-defined significance level alpha in cell **L43** with the calculated p-value in cell **L45**. If the p-value is less than alpha, the formula returns "We reject the null hypothesis", indicating statistical significance between the groups. If the p-value is greater than or equal to alpha, the formula returns **"We fail to reject the null hypothesis",** suggesting there is not enough evidence to conclude the groups are different. This formula simplifies the interpretation of T-test results by providing a straightforward conclusion based on the p-value and significance level.
To calculate the difference in means between two groups, use the formula **L52-L51** in a spreadsheet. This will give you the point estimate

needed to construct a 95% Confidence Interval (CI) using a T-distribution.

To find the product of the values in cells **L53 a**nd **L54** in a spreadsheet, use the formula **=L53×L54.** This can be helpful when constructing a 95% Confidence Interval **(CI)** using the T-distribution because it allows you to calculate the **Margin of Error (MOE)**.

Code For The Novelty Effect

```
-- Start the SQL query by selecting the columns we need
SELECT
    a1.dt,                          -- Date of the activity
    g1.group as test_group,             -- Group to which the user belongs
   COUNT(DISTINCT a1.uid) as total_users,  -- Count of unique users
    SUM(ROUND(a1.spent,2)) as total_purchase, -- Sum of all
purchases, rounded to 2 decimal places
    AVG(ROUND(a1.spent,2)) as average_purchase  -- age purchase per
us  Averer, rounded to 2 decimal places
-- From the 'activity' table, aliased as 'a1'
FROM
    activity a1
-- Perform a JOIN operation with the 'groups' table, aliased as 'g1'
-- Matching is done based on the 'uid' column in both tables
JOIN
    groups g1
    ON a1.uid = g1.uid
```

```
-- Filter the rows based on the date range, from '2023-01-25' to
'2023-02-06'
WHERE
    a1.dt BETWEEN  '2023-01-25' AND '2023-02-06'
-- Group the result set by date and group
GROUP BY
    a1.dt,
    g1.group

-- Sort the result set first by date in descending order, then by group
ORDER BY
    a1.dt DESC,
    g1.group;
```