In [64]:

```
import pandas as pd
import numpy as np
import scipy.stats as stats
```

Data Loading

In [65]:

```
# Importing campaign data and remove "Control" mailer type
campaign_data = pd.read_excel("grocery_database.xlsx", sheet_name= "campaign_data")

campaign_data = campaign_data[campaign_data['mailer_type'] != "Control"]
print(campaign_data.head())
```

```
   customer_id  campaign_name campaign_date mailer_type  signup_flag
0           74  delivery_club    2020-07-01     Mailer1            1
1          524  delivery_club    2020-07-01     Mailer1            1
2          607  delivery_club    2020-07-01     Mailer2            1
3          343  delivery_club    2020-07-01     Mailer1            0
4          322  delivery_club    2020-07-01     Mailer2            1
```

Calculating signup rates

In [66]:

```
#Create contingency table
contingency_table = campaign_data.groupby(by='mailer_type')['signup_flag'].value_counts(

# Row sum, column sums and signup rates
row_sum = contingency_table[0] + contingency_table[1]
column_sum = contingency_table.iloc[0] + contingency_table.iloc[1]
signup_rates = contingency_table[1]/row_sum
total_observations = sum(row_sum)

print(contingency_table)
print("\nMailer1 signup rate:", (signup_rates[0]))
print("Mailer2 signup rate:", (signup_rates[1]))
```

```
signup_flag    0    1
mailer_type
Mailer1      252  123
Mailer2      209  127

Mailer1 signup rate: 0.328
Mailer2 signup rate: 0.37797619047619047
```

Chisquare test of independence

Null Hypothesis (H_0): There is no significant relationship between mailer_type and signup_flag.

Alternate Hypothesis (H_1): There is a significant relationship between mailer_type and signup_flag.

In [67]:

```
# Using scipy.stats.chi2_contingency
stat, p, dof, expected = stats.chi2_contingency(contingency_table)

print("Chisquare critical value: ", stats.chi2.ppf(0.95, df=dof))
print("Chisquare observed value: ", stat)
print("Chisquare p-value: ",p)
```

```
Chisquare critical value:  3.841458820694124
Chisquare observed value:  1.728424144871394
Chisquare p-value:  0.1886122739808747
```

Conclusion

Since the p-value is 0.18 which is greater than 0.05 or chisquare value is less than critical value, we can not reject the null hypothesis. Hence, there is no significant relationship between mailer_type and signup_flag.

Bonus: Why are we using chi-square distribution in this case?

Reasons:

1. There are 1 (or more) categorical variables(like mailer_type).
2. All the observations in the given dataset are independent of each other.
3. The given set of categorical variables in the dataset are mutually exclusive.

Gaussian Distribution

1. Before we use Gaussian distribution, we need to check if the data is actually following normal distibution.
2. Gaussian distribution is typically used to compare a variable with a value. (like when a variable is compared to population mean) For this dataset, we are comparing 2 different variables (mailer1 and mailer2).