Data Analysis and Statistical Inference

Introduction

To stimulate the reactivation of dormant prepaid subscribers on its network (i.e. reduce churn), a leading south American mobile Operator (in Brazil) uses broadcast SMS to advertise attractive offers to dormant users. Similar SMS offers are also sent to active users to stimulate more frequent and larger top-ups. As is the case in most emerging markets, this Operator's user base is heavily skewed towards prepaid usage. Therefore, the slightest improvement in churn rates has a tremendous impact on this MNO's profitability and overall financial performance.

The Operator retained Persado's (www.persado.com) solution for MNOs. Specifically, Persado was hired to generate SMS messages that drive the highest dormant reactivation rate, increase top-up frequency rate and amounts.

Henceforth the group of mobile subscribers that received Persado's SMS messages will be referred to as "Persado" group. The "control" group of our study consists of mobile subscribers that did not receive any "promo" SMS message; henceforth the group of such users will be referred to as "Baseline" group.

This study will evaluate whether the proportion of "top-up" in "Persado" group is different than that in "Baseline" (control) group.

Data

In cooperation with the mobile Operator Company (OpCo), a study designed to test the "top-up" rate under "treatment" or no "treatment" conditions. A group of mobile subscribers, randomly assigned, to receive messages generated by "Persado". Another group of mobile subscribers did not receive any "promo" SMS message but their "top-up" rate also monitored. Observation period for this study (experiment) was one (1) week.

The data set collected consists of:

- **region**: i.e. two states in Brazil: "Minas Gerais" & "Rio de Janeiro". Each mobile subscriber of this study lives in either of these two regions.
- **mask_msisdn**: this is the "masked" msisdn (phone number) of each of the mobile subscribers in this study.
- **msg_source**: two values; either "PERSADO" or "BASELINE". In the former case, mobile subscriber received an SMS message generated by "Persado". In the latter case, he/she did not receive a message.
- **top_up**: two values; either "0" or "1". In the former case, mobile subscriber did not make a "top-up" (i.e. No). In the latter case, he/she made a "top-up" (i.e. Yes).

In practice, the study consists of mobile subscribers living in either of the above mentioned states in Brazil. About 95% of those mobile subscribers received a "top-up promo" message ("Persado" group) and the rest 5% did not receive a message ("Control/Baseline" group; its size is sufficient for the purposes of our study).

The scope of the study was to evaluate generally whether there is a relationship between "top-up" [a] and "promo" message received ([b]; "msg_source") or "region" [c]. So, we have three variables in our data set:

- [a] top-up "Yes" or "No".
- [b] receive message or not.
- [c] region "A" vs. region "B".

In theory, we can consider the relationship between only two of the above mentioned variables without considering any other possible confounders. That is, we can evaluate the relationship between [a] and [b] or [a] and [c]. For the purposes of this study (experiment), we evaluated the relationship between "top-up" [a] and "promo" message received or not [b]. Specifically:

- **top_up**: two values; either "0" or "1". Type: "categorical".
- **msg_source**: two values; either "PERSADO" or "BASELINE". Type: "categorical".

The data set contains "393,334" data cases of mobile subscribers (units of experiment), who either received or not a "top-up promo" SMS message and made or not a "top-up" finally.

The type of the conducted study is "experiment", with the following characteristics briefly:

- fair experiment; random sample.
- same sample proportions for each region (state).
- same sample "split" into "Persado" and "Control/Baseline" groups within each region (state).

In some more detail and according to the principles of "experimental design":

- **control**: we compared treatment of interest (i.e. broadcast of "top-up promo" SMS messages generated by "Persado") to a control group ("Baseline" group that received no messages).
- **randomize**: we randomly assigned subjects (i.e. mobile subscribers) to treatment and control groups ("Persado" or "Baseline" groups respectively; i.e. receiving or not a "promo" message). So, we tried to equalize the groups with respect to any confounding variables and with each subject's (subscriber) chances of ending up in one condition or the other ("top-up" or not) being independent of the chances of other subjects. In this way, any difference in the response (top_up) should be attributable to the explanatory variable (msg_source). Thus, we tried to establish causal connection between the explanatory (i.e. receive or not a "promo" message) and response (i.e. make or not a "top-up) variable.
- **replicate**: within this study, we replicated by collecting a sufficiently large sample.

- **block**: we suspected that the "promo" SMS messages might affect differently mobile subscribers living in different "country states". So, we first grouped subjects (i.e. mobile subscribers) into blocks (depending on the "state" they live in). Then we randomized cases (units of experiment, i.e. mobile subscribers) within each block (i.e. "state") to treatment groups ("Persado" or "Baseline" groups; i.e. receiving or not a "promo" message).

Actually during the "sampling" phase, we applied the "cluster" method: we divided the population clusters (i.e. country states), randomly sampling a two of those clusters (states), and then randomly sampling from within these two clusters.

After a set amount of time (observation period one week), we recorded the amount of "top-ups" for each mobile subscriber in this study.

Eventually, we conducted a hypothesis test to evaluate whether any relationship (causality) we find is indeed statistically significant.

The population of interest of this study is (prepaid) mobile subscribers of the OpCo. By applying a random sampling method (cluster), our results can be generalized to this population. At the same time, by conducting an "experiment" with a "random assignment", causal conclusions can be made.

As a last note, a potential source of bias might be worth mentioning: the percentage of mobile subscribers, who actually received "successfully" a "top-up promo" SMS message may not necessarily constitute a representative sample from the whole population. In such a case, it is possible that while a "significant" percentage of the subscribers (who received a "promo" message) made finally a "top-up", this may not hold true for those subscribers who did not received successfully such a message (due to e.g. network related problems; although being members of "Persado" group). However, the fraction (of those randomly sampled users), who did not receive successfully a "promo" message, should be assumed as "random" and too low (a matter of few thousands) in relation to the total number of mobile subscribers of the "Persado"(treatment) group (a matter of millions).

Exploratory Data Analysis

The "explanatory data analysis" suggests that possibly there is a statistically significant difference between the "top-up" rates in "Persado" and "Baseline" groups (or alternatively there is a statistically significant association - causality - between "top-up" rate and "promo" messages generated by "Persado"). Nevertheless, this difference in sample statistics, suggesting dependence between the two "categorical" variables, may be due to chance; thus the statistical significance of that difference will be precisely evaluated through a (two-sided) hypothesis testing that shall be presented in the following paragraph.

Below you can find the conducted "explanatory data analysis" presenting statistics and plots produced by "R" for the uploaded data set.

Summary statistics

```
# Contingency Table of two 'Categorical' Variables (2 x 2).

msg_source_vs_top_up = table(top_up, msg_source)
msg_source_vs_top_up
```

```
##       msg_source
## top_up BASELINE PERSADO
##      0    15480  283445
##      1     4531   90878
```

```
# 'Row' Marginal Totals - 'Response' Variable.

margin.table(msg_source_vs_top_up, 1)
```

```
## top_up
##      0      1
## 298925  95409
```

```
# 'Column' Marginal Totals - 'Explanatory' Variable.

margin.table(msg_source_vs_top_up, 2)
```

```
## msg_source
## BASELINE  PERSADO
##    20011   374323
```

```
# 2-Way Frequency Table - 'Cell' Percentages.

prop.table(msg_source_vs_top_up)
```

```
##       msg_source
## top_up BASELINE PERSADO
##      0  0.03926 0.71879
##      1  0.01149 0.23046
```

```
# 2-Way Frequency Table - 'Row' Percentages of 'Response' Variable.

prop.table(msg_source_vs_top_up, 1)
```

```
##       msg_source
## top_up BASELINE PERSADO
##      0  0.05179 0.94821
##      1  0.04749 0.95251
```

```
# 2-Way Frequency Table - 'Column' Percentages of 'Explanatory' Variable.

prop.table(msg_source_vs_top_up, 2)
```

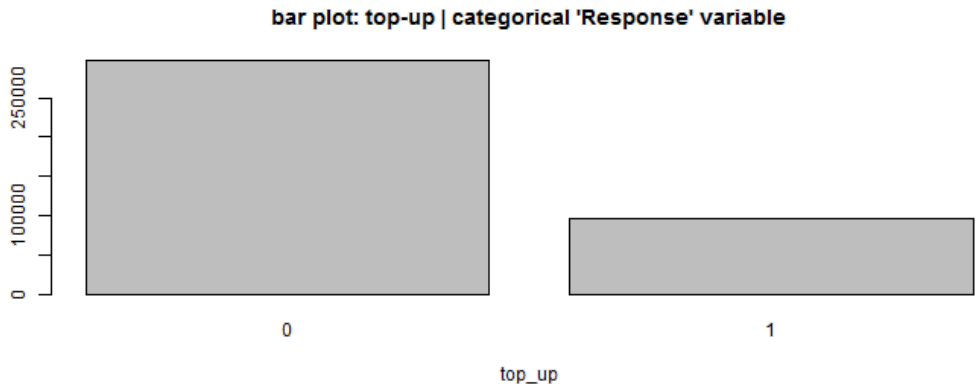
```
##       msg_source
## top_up BASELINE PERSADO
##      0   0.7736  0.7572
##      1   0.2264  0.2428
```

Recall that the hypothesis test will focus on the difference between the “top-up” rates under “treatment” or no “treatment”. So in the case of our sample, we have calculated:

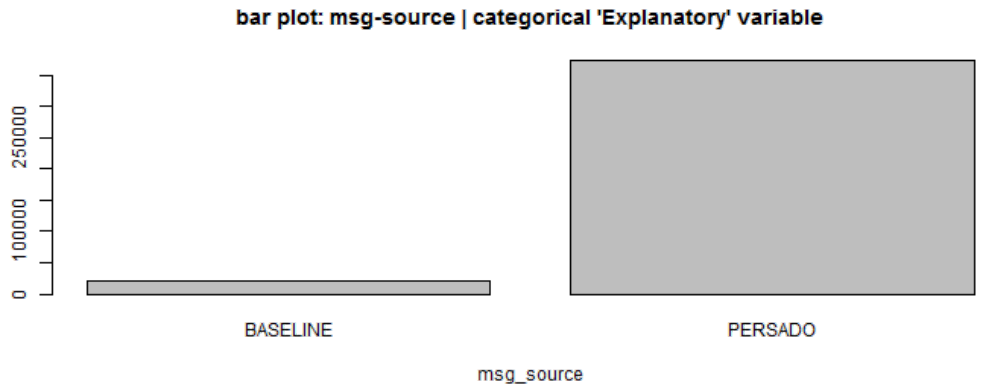
- treatment group - “Persado” top-up rate: $\approx 24.28\%$(relative frequency - point estimate).
- no treatment (control) group - “Baseline” top-up rate: $\approx 22.64\%$(relative frequency - point estimate).

Visualization

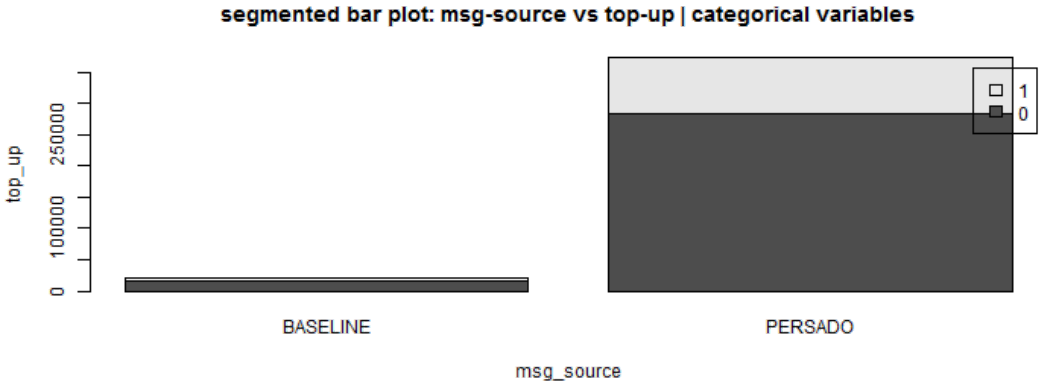
```
# Bar Plot of 'Response' Variable.

barplot(table(top_up), main = "bar plot: top-up | categorical 'Response' variable",
    xlab = "top_up")
```



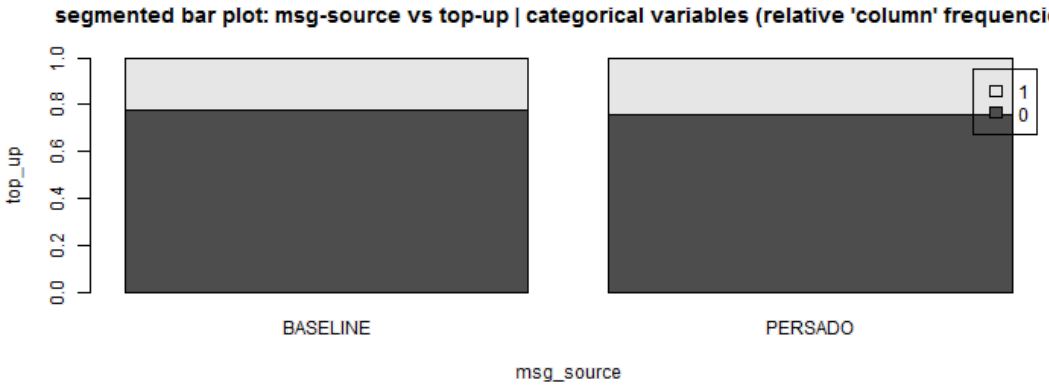

```
# Bar Plot of 'Explanatory' variable.

barplot(table(msg_source), main = "bar plot: msg-source | categorical 'Explanatory'
variable",
    xlab = "msg_source")
```




```
# Stacked Bar Plot of two Categorical Variables.

barplot(msg_source_vs_top_up, main = "segmented bar plot: msg-source vs top-up |
categorical variables",
    xlab = "msg_source", ylab = "top_up", beside = FALSE, legend =
rownames(msg_source_vs_top_up))
```

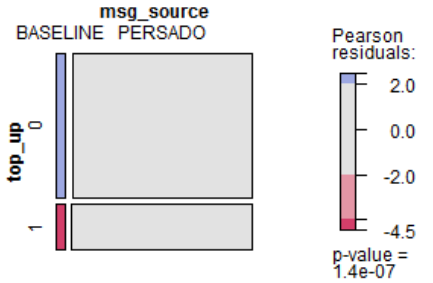



```
# Stacked Bar Plot of two Categorical Variables - Relative 'Column'
# Frequencies.

barplot(prop.table(msg_source_vs_top_up, 2), main = "segmented bar plot: msg-source vs
top-up | categorical variables (relative 'column' frequencies)",
    xlab = "msg_source", ylab = "top_up", beside = FALSE, legend =
rownames(msg_source_vs_top_up))
```




```
# Mosaic Plot of two Categorical Variables.

mosaic(msg_source_vs_top_up, shade = TRUE, legend = TRUE, main = "mosaic plot: msg-source
vs top-up | categorical variables")
```

mosaic plot: msg-source vs top-up | categorical variables



Inference

A hypothesis test shall be conducted to evaluate if the data provide strong evidence that the proportion of “Persado top-ups” is different than the proportion of “Baseline top-ups” (the control group). In other words, this test shall evaluate whether there is a statistically significant association - causality - between “top-up” rate and “promo” messages generated by “Persado”.

Let p_p and p_B represent the proportion of “top-up” in "Persado” and “Baseline” groups respectively. The “null” and “alternative” hypotheses can be stated as follows:

- H_0 : the proportion of “top-up” is the same in “Persado” and “Baseline” groups, $p_P = p_B$
- H_A : the proportion of “top-up” is different in “Persado” and “Baseline” groups, $p_P \neq p_B$

“Random assignment” was used, so the observations in each group are independent. On the other hand, the mobile subscribers in the study must be assumed as representative of those in the general population. So, we can also confidently generalize the findings to the population. Each group is a “random sample” less than 10% of the whole population. In this way, sampled members of “Persado” group are independent of each other; sampled members of “Baseline” group are as well. Thus, **“independence” condition is satisfied**.

In addition, let X_B and X_P represent the number of “successes” (top-up) in “Baseline” and “Persado” sampled groups respectively. Let also n_B and n_P represent the sample sizes of “Baseline” and “Persado” groups respectively. The **“success-failure” condition**, which we shall check using the “pooled” proportion:

$$\hat{p_{pool}} = \frac{(X_B + X_P)}{(n_B + n_P)} = \frac{(4531 + 90878)}{(20011 + 374323)} \approx 0.2419$$

is satisfied for both “Baseline”:

- $n_B * \hat{p_{pool}} = 20011 * 0.2419 = 4840.6609 \geq 10$
- $n_B * (1 - \hat{p_{pool}}) = 20011 * 0.7581 = 15170.3391 \geq 10$

and “Persado” groups:

- $n_P * \hat{p_{pool}} = 374323 * 0.2419 = 90548.7337 \geq 10$
- $n_P * (1 - \hat{p_{pool}}) = 374323 * 0.7581 = 283774.2663 \geq 10$

With the conditions met, we can finally assume that **the sampling distribution of the difference between the two proportions is nearly normal** (Central Limit Theorem) and we shall **conduct a “two-sided” hypothesis test, at 5% significance level,** evaluating if “Persado” and “Baseline” group members are equally likely to make a “top-up”, i.e evaluating the “null” (and “alternative”) hypotheses as stated above. Hence:

- $H_0 : p_P - p_B = 0$ and $H_A : p_P - p_B \neq 0$
- $(\hat{p}_P - \hat{p}_B) \sim N(mean = 0, SE = \sqrt{\frac{\hat{p_{pool}} * (1 - \hat{p_{pool}})}{n_B} + \frac{\hat{p_{pool}} * (1 - \hat{p_{pool}})}{n_P}} = \sqrt{\frac{0.2419 * 0.7581}{20011} + \frac{0.2419 * 0.7581}{374323}} \approx 0.0031)$
- point estimate: $\hat{p}_P - \hat{p}_B = 0.2428 - 0.2264 = 0.0164$

Recall that:

- $\hat{p}_B = \frac{4531}{20011} \approx 0.2264$
- $\hat{p}_P = \frac{90878}{374323} \approx 0.2428$

are sample “top-up” proportions taken from samples of size:

- $n_B = 20011$
- $n_P = 374323$

for “Baseline” and “Persado” groups respectively.

The “Test” statistic

$$z = \frac{(\hat{p}_P - \hat{p}_B) - 0}{SE} = \frac{0.0164}{0.0031} = 5.2903$$

follows the standard normal distribution (with mean = 0 and standard deviation = 1) and is used to compute the “p-value” for the standard normal distribution; the probability that a value at least as extreme as the test statistic would be observed under the “null” hypothesis [P(observed or more extreme test statistic | H_0 true)]. Given the “null” hypothesis that the population proportions are equal, the “p-value” for testing H_0 against H_A as stated above (i.e. a “two-sided” hypothesis) is:

p-value = $2 * P(Z > |z|) = 2 * P(Z > 5.2903) = 2 * (1 - P(Z \leq 5.2903)) \approx 0$

```
p_value = 2 * pnorm(5.2903, lower.tail = FALSE)
p_value
```

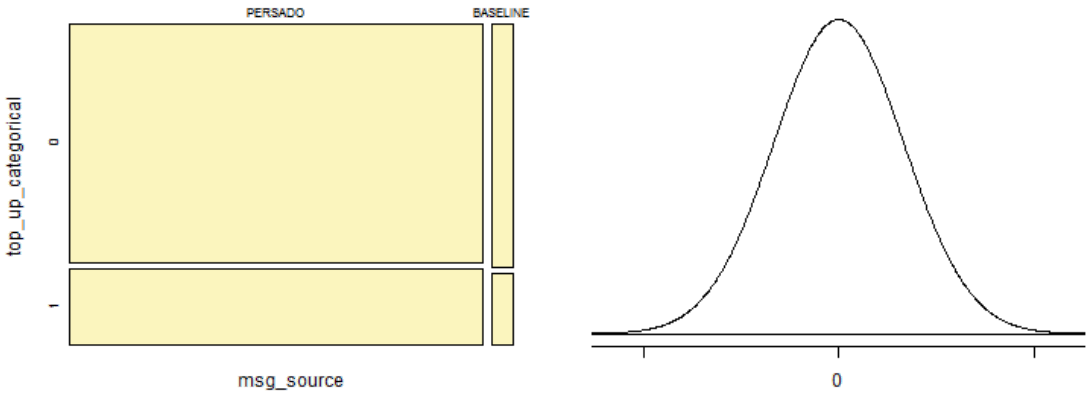
```
## [1] 1.221e-07
```

This is **statistically significant at the “5%” level**; “p-value”“ is less than "0.05”. Thus, we **reject “null” hypothesis H_0. The data provide convincing evidence that the proportion of “top-up” in “Persado” group is different than that in “Baseline” (control) group. The data also indicate that a higher proportion of “Persado” members make “top-up” than that of “Baseline” members (i.e. the data indicate the direction after we reject H_0).**

```
source("http://bit.ly/dasi_inference")

top_up_categorical = factor(top_up, levels = c("0", "1"))

inference(y = top_up_categorical, x = msg_source, est = "proportion", type = "ht",
    method = "theoretical", null = 0, alternative = "twosided", success = "1",
    order = c("PERSADO", "BASELINE"), conflevel = 0.95, siglevel = 0.05, sum_stats = TRUE,
    eda_plot = TRUE, inf_plot = TRUE, inf_lines = TRUE)
```

```
## Response variable: categorical, Explanatory variable: categorical
## Difference between two proportions -- success: 1
## Summary statistics:
##       x
## y      PERSADO BASELINE    Sum
##   0    283445    15480 298925
##   1     90878     4531  95409
##   Sum  374323    20011 394334
```

```
## Observed difference between proportions (PERSADO-BASELINE) = 0.0164
## H0: p_PERSADO - p_BASELINE = 0
## HA: p_PERSADO - p_BASELINE != 0
## Pooled proportion = 0.2419
## Check conditions:
##    PERSADO : number of expected successes = 90567 ; number of expected failures =
283756
##    BASELINE : number of expected successes = 4842 ; number of expected failures = 15169
## Standard error = 0.003
## Test statistic: Z =  5.263
## p-value =  0
```



We shall also estimate the difference between the two proportions. That is, the “parameter of interest” is the difference between the “top-up” proportions of all “Persado” group members and all “Baseline” ones: $p_P - p_B$. While the “point estimate” is the difference between the “top-up” proportions of sampled “Persado” group members and sampled “Baseline” ones: $\hat{p}_P - \hat{p}_B$.

Recall that the **"independence" condition within and between sampled groups is satisfied** (as stated above; i.e. random assignment & random sample from less than 10% of the whole population for each group). Also, **each sample meets the "success-failure" condition:**

- Baseline: 4531 successes ≥ 10 and 15480 failures ≥ 10
- Persado: 90878 successes ≥ 10 and 283445 failures ≥ 10

So, we can again assume that **the sampling distribution of the difference between the two proportions is nearly normal** (Central Limit Theorem).

Let z^* represent the upper "critical" value from the standard normal distribution; in our case $z^* = 1.96$, assuming a 95% level "confidence interval". Let also SE represent the "standard error" for the difference between two proportions, for calculating a confidence interval, which is computed as follows:

$$SE = \sqrt{\frac{\hat{p}_B * (1-\hat{p}_B)}{n_B} + \frac{\hat{p}_P * (1-\hat{p}_P)}{n_P}} = \sqrt{\frac{0.2264 * 0.7736}{20011} + \frac{0.2428 * 0.7572}{374323}} \approx 0.0030$$

Recall again that:

- $\hat{p}_B = \frac{4531}{20011} \approx 0.2264$
- $\hat{p}_P = \frac{90878}{374323} \approx 0.2428$

are sample "top-up" proportions taken from samples of size:

- $n_B = 20011$
- $n_P = 374323$

for "Baseline" and "Persado" groups respectively.

An approximate **95% level "confidence interval" for the difference between the two proportions** $p_P - p_B$ is:

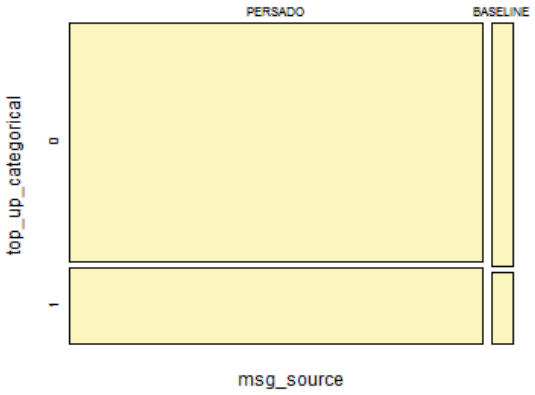
$$(\hat{p}_P - \hat{p}_B) \pm z^* SE = (0.2428 - 0.2264) \pm 1.96 * 0.0030 = (0.0105, 0.0223)$$

We are 95% confident that the true "top-up" proportion of "Persado" group is 1.05% to 2.23% higher than the true "top-up" proportion of "Baseline" group. Alternatively, we can assume that 95% of random samples will produce 95% confidence intervals that include the true difference between the population "top-up" proportions of "Persado" and "Baseline" groups.

Based on the 95% level "confidence interval" we calculated, we can indeed **expect to find a significant difference (at the equivalent 5% "significance level") between the population "top-up" proportions of "Persado" and "Baseline" groups.** Recall that the "null" hypothesis stated as $H_0 : p_P - p_B = 0$. So based on the just computed interval, we must reject H_0; the confidence interval does not contain "0".

```
inference(y = top_up_categorical, x = msg_source, est = "proportion", type = "ci",
    method = "theoretical", success = "1", order = c("PERSADO", "BASELINE"),
    conflevel = 0.95, sum_stats = TRUE, eda_plot = TRUE, inf_plot = TRUE, inf_lines =
TRUE)
```

```
## Response variable: categorical, Explanatory variable: categorical
## Difference between two proportions -- success: 1
## Summary statistics:
##       x
## y      PERSADO BASELINE    Sum
##   0    283445    15480 298925
##   1     90878     4531  95409
##   Sum  374323    20011 394334
```



```
## Observed difference between proportions (PERSADO-BASELINE) = 0.0164
## Check conditions:
##    PERSADO : number of successes = 90878 ; number of failures = 283445
##    BASELINE : number of successes = 4531 ; number of failures = 15480
## Standard error = 0.003
## 95 % Confidence interval = ( 0.0104 , 0.0223 )
```

Conclusion

We designed an “experimental” study, utilizing a treatment (Persado) and a control (Baseline) group, and collecting a data set that consisted of the following (but not limited to) “categorical” variables: “top_up” (yes/no) & “msg_source” (PERSADO/BASELINE).

A “two-sided” hypothesis test conducted investigating whether the data provide convincing evidence that “top-up” and “msg_source” variables are independent (null hypothesis - H_0) or dependent (alternative hypothesis - H_A). In other words, H_0 assumes that no effect on “top-up” due to “msg_source” observed, while H_A assumes that there is a “msg_source” effect. Thus, we evaluated whether the observed difference in “top-up” rates is simply by chance or not.

The outcome of the hypothesis test, at a 5% “significance level”, reveals that the data provide convincing evidence that the proportion of “top-up” in “Persado” group is different than that in “Baseline” (control) group, rejecting “null” hypothesis - H_0.

We also estimated the difference between the two proportions, for a 95% level “confidence interval”, confirming the rejection of the “null” hypothesis - H_0 as well as stating that we are **95% confident that the true “top-up” proportion of “Persado” group is about 1% to 2% higher than the true “top-up” proportion of “Baseline” group.**

Thus, the results of both methods agree with each other.

As a last note, we should further investigate the effect of other demographic variables on the “top-up” rate, such as (but not limited to) average age, income, region of residence, marital status, etc.

References

Persado (www.persado.com). PNGM (Prepaid Next Generation Marketing) campaign data. Week: 6, February 2014 (observation period 1 week). The data is courtesy of Persado (not available on-line).

Appendix

The data set used in this study contains data similar to the following. The first row consists of the data headers. In the data set file uploaded, the columns, in each of those cases, are being delimited by a “semicolon” character.

```
head(format(campaign, justify = "right", width = 12, scientific = FALSE), n = 20)
```

```
##                  region  mask_msisdn   msg_source       top_up
## 1    BRAZIL Minas Gerais   3289006301      PERSADO            0
## 2    BRAZIL Minas Gerais   3183003983      PERSADO            1
## 3    BRAZIL Minas Gerais   3192125465      PERSADO            0
## 4    BRAZIL Minas Gerais   3589163961      PERSADO            0
## 5    BRAZIL Minas Gerais   3489305436     BASELINE            0
## 6  BRAZIL Rio de Janeiro   9192783383      PERSADO            1
## 7    BRAZIL Minas Gerais   3184366595     BASELINE            0
## 8    BRAZIL Minas Gerais   3193970477     BASELINE            1
## 9    BRAZIL Minas Gerais   3186257577      PERSADO            0
## 10 BRAZIL Rio de Janeiro   9184779186      PERSADO            0
## 11   BRAZIL Minas Gerais   3590756165     BASELINE            0
## 12   BRAZIL Minas Gerais   3387771593      PERSADO            0
## 13 BRAZIL Rio de Janeiro   9192939630      PERSADO            0
## 14 BRAZIL Rio de Janeiro   9990171350      PERSADO            0
## 15   BRAZIL Minas Gerais   3289142792     BASELINE            0
## 16 BRAZIL Rio de Janeiro   9292072202      PERSADO            0
## 17   BRAZIL Minas Gerais   3184758258      PERSADO            1
## 18   BRAZIL Minas Gerais   3192790719      PERSADO            1
## 19 BRAZIL Rio de Janeiro   9184931576      PERSADO            0
## 20 BRAZIL Rio de Janeiro   9188997223     BASELINE            0
```