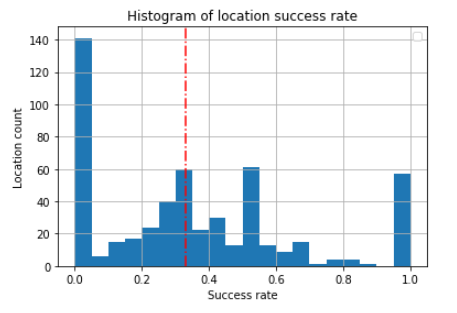
To Do:

- Deal with categorical variables.

- Decide whether to make groups inside local, departm and stop\_reason.

- Build pipeline.

How to decide whether the success rate associated with a given feature’s category is meaningful, i.e., whether this category being present really changes the probability of success?



In the example above we have grouped observations by all the different locations where searches have taken place and plotted the histogram of the success rate. We see that some locations have a success rate higher than average (on the right of the red line) and other locations have a success rate below average. Some locations even have a success rate of 100% and others oh 0%. In order to decide whether the observed success rate is meaningful or is just pure chance, we perform a binomial test with the null hypothesis being.

# Client requirements

## Summary

## Requirements clarifications

*Awkward Problem Solutions™* has been asked to comply with the following requirements:

1. A minimum 50% success rate for searches (when a car is searched, it should be at least 50% likely that contraband is found)
2. No police sub-department should have a discrepancy bigger than 5% between the search success rate between protected classes (race, ethnicity, gender)
3. The largest possible amount of contraband found, given the constraints above.

The third requirement is the most evident. If we can increase the percentage of contraband found we also increase the certainty of punishment, producing a stronger deterrent effect which prevents more people from committing contraband crimes. From a specification perspective, this is equivalent to ask for a maximization of the model’s recall score: for a given number of stops the existing contraband either is searched and found (true positives) or not searched and not found (false negative), and we want to find as much contraband as possible (true positives) given its total amount (true positives + false negative).

Regarding the first requirement, we assume the existence of a cost associated with the searching procedure and a benefit associated with finding contraband. The rationale for this requirement should be the fact that the client values the benefit resulting from finding contraband (let’s call it 2*s*) as twice the cost of searching (call it *s*), on average. In this way, since every time a search is made we incur in the cost *s*, searching is ‘cost-effective’ only if the probability of finding contraband is, at least, 50%, i.e., if we succeed in finding contraband in 50% of the searches performed. This requirement defines the probability threshold to be used in theclassification.

the rational for The client values the success and the search with a cost, the implicit idea in this requirement is that

Loss / benefit

# Dataset analysis

## General analysis

Missing values; unique(), .nunique(), .value\_counts(); max, idxmax, min, idxmin, mean, median, mode

Histrograms, distributions and plots.

Variance, covariance, correlation

## Business questions analysis

## Conclusions and Recommendations

# Modeling

## Model expected outcomes overview

## Model specifications

## Analysis of expected outcomes based on the training set

In-sample-error (ISE) or training error vs. Out-of-sample error (OSE) or testing error

## Alternatives considered

## Known issues and risks

# Model Deployment

## Deployment specifications

## Known issues and risks

## Dataset technical analysis

# Annexes

## Business questions technical support

## Model technical analysis