Instacart Shopping Data

Inferential Statistics

Sava Dashev

The dataset has variables depicting the day and time of the order, the aisle, the department from which the product is from, and the number of days since prior order.

We can investigate the reorder ratio by day of week, hour of day, aisle and department.

The dataset is large. For each value of the variable there are thousands, and in some instances ten thousands or hundred thousands of products in each category. When the sets are that large, even very small differences in proportion are statistically significant. The reorder ratio is the success rate of a Bernoulli distribution for each particular product and the number of the ordered is the number of trials, and the number of reorders is the number of successes.

There are still products with very small number of orders and reorder rate that would not allow analysis. We look at products that have more than 50 products and we need at least 5 reorders, or at least 5 failures (not reorders) to be able to make sound statistical analysis.

We compare the *z-*value of the pooled set to the predetermined z-value. We choose level 0.05 for our tests.

For two products with proportions $p\_1$ and $p\_2$, we calculate pooled proportion by adding the success (x1 + x2), and divide by pooled population (n1 + n2).

Since this is a Bernoulli distribution, the pooled error is

We calculate z value of the difference of the observed proportion and using the error calculated for the pooled population to calculate the z-value of the difference. If the z-value is bigger than the predetermined z-value calculated level of significance (we use 0.05), then the null hypothesis is rejected.

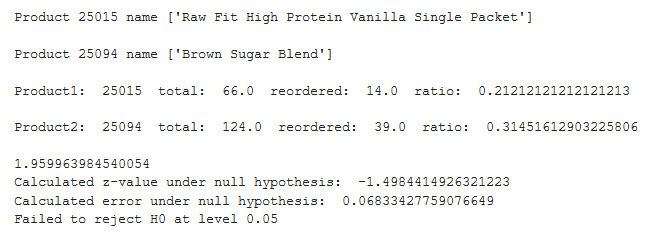
We implement a function check\_hypothesis to calculate the calculations with two different sets of input parameters.

We look at different products. There are almost 50,000 products, and even when we restrict the minimum number of products ordered to 50, we are left with about 38,000 products.

Hypothesis: The differences between the reordered proportion are statistically significant if we have the difference in reorder proportion no less than 0.15.

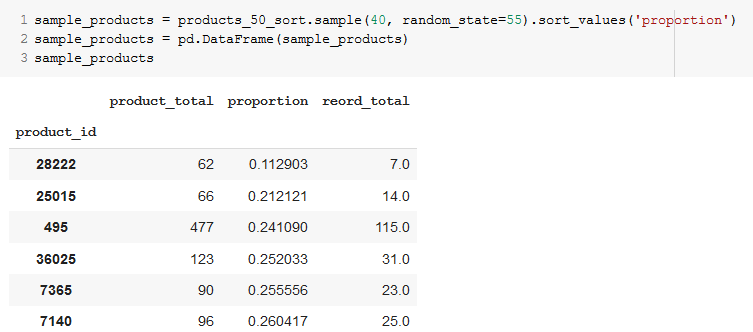
We checked difference of reorder proportions 0.1. One out of 4 hypothesis was not rejected at that level. We check at level 0.15, using different products from the same random set.

Hypothesis check with difference of reorder proportion 0.1:

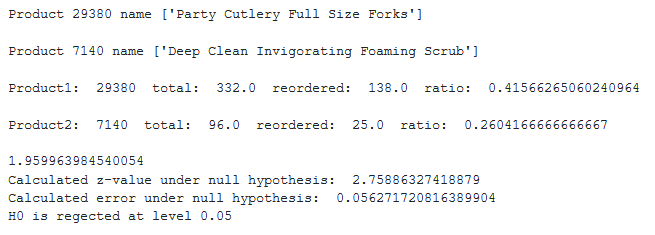


We checked if lower values of the difference is statistically significant, but in some cases, we failed to reject the null hypothesis.

We choose a random sample of products and choose products with the desired difference between reorder proportion. We used products with small sample size. With product with larger sample size, the statistical significance will increase. Here is part of that list:

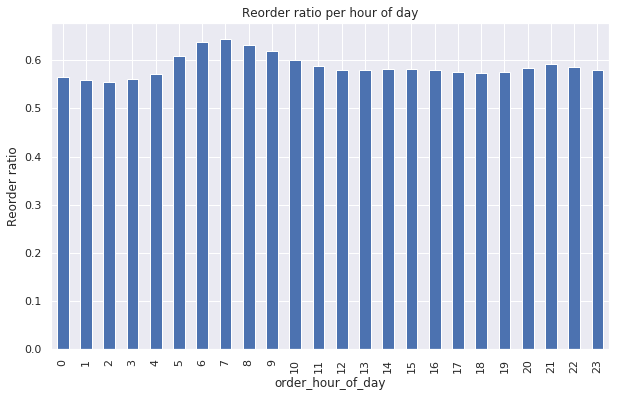


We used four pairs of the random sample list that have that property. One of the checked is shown below.

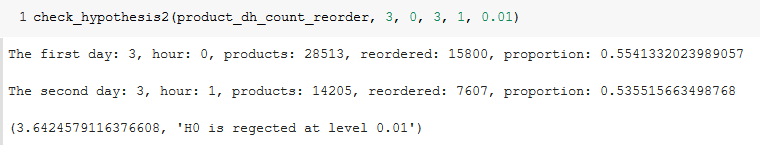


There is a pair of variables that describe day and hour of making an order. We can investigate if the reorder ratio varies with those.

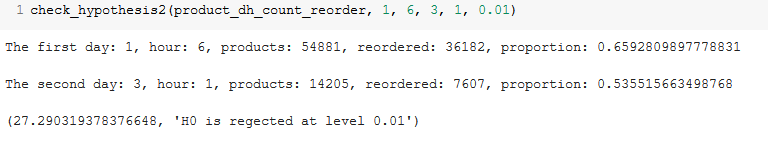
And reorder ratio by hour of day:



Hypothesis: The differences in reorder ratio proportions for products ordered in different days and times is statistically significant, when the difference in ratio is more than 0.02. In a few cases, we fail to reject the above hypothesis, but in general, the reorder proportions for different days and times belong to different populations.



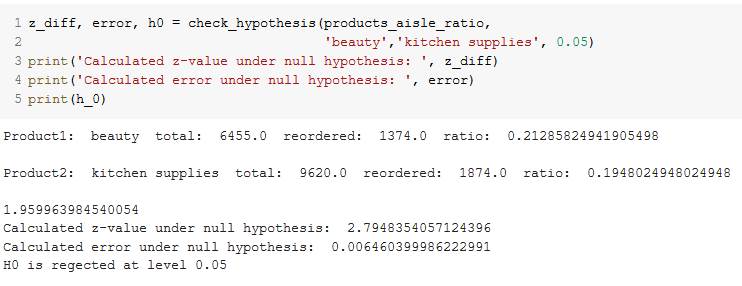
To check the hypothesis, we use a version of check\_hypothesis function, designed to test two different days and times. When differences ratio is about 0.02, we may fail to reject the hypothesis. For most cases, the differences may be more than 0.02, or the sets of products are large, we are able to reject the null hypothesis that two populations are samples of one and the same population.



The z-value shown above is 27.

We can form hypothesis about reorder ratio in different aisles and different departments. Each aisle form separate population. The null hypothesis will be The proportion ratios of different departments are samples of the same population, i.e. are equal, the shown difference is due to chance. The alternative hypothesis is that the reorder proportions are different. We check the hypothesis at level 0.05.

Using the same function, we can streamline the check. We choose aisles with small number of items sold and close proportion. If the hypothesis is rejected for these, it is much more likely to be rejected for aisles with larger difference of proportion difference, or larger number of items sold and/or reordered.



Departments

The department reorder ratios vary from 0.32 to 0.67. We can check two of the departments with close ratios and smaller population to check if they are statstically different.

