Exploring diffrences in Satellite's lifetime using permutation tests

Netta Shafir and Shaked Amar

August 2022

Introduction

During the cold war, one of the centers of the struggle between the United States and the Soviet Union was known as the **"Space Race"** - the try of the two rivals to achive a supermacy in spaceflight capabilities. Since then, the presence of a country in the outer space is considered as main indicator for it's strength.

Today, a lot of countries has satellites in space for varied applications. Nevertheless, the majority of the satellites is still belong to few, and we were interested in the differences of the sattelites between those countries. In particular, we would like to test the hypotheses that there is a difference in the expected value lifetime of satellites of those countries. Under several assumptions, those kind of hypotheses can be tested within F-test as part of ANOVA (anylasis of variance), which are a generalization of the t-test in the case of more than two groups. Notwithstanding, an important assumption of the t-test and the F-test in our context of testing the hypotheses that two of more groups have the same mean, is that the response variable is Normaly distributed. As we'll see soon enough, this assumption doesn't hold in our case, and therefor we will conduct similar tests, which are not-parametric and therefor are free from any assumption about the distribuition of the data. These tests also involves randomness.

The dataset

The dataset we investigae has been taken from Kaggle[1]. It contains data about all the satellites that launched since 1957 - The year in which the first satellite, Sputnik 1, was sent to space by the USSR. The variable **owner** refers the country that holds the satellite. The variables **launch_date** and **flight_ended** refer the dates in which the satellite was sent to space and got out of use, respectively. As for the **status** feature, it tells the current activity status of the satellite[2], so in particular a satellite whose status is **Operational** should not have any value in the **flight_ended** field. Therefor, we were intrested only in satellites their status is **decayed**, about which we could get the full number of days until out of use.

First exploration of the data

To minimize time-related influences, we focus on satelittes that launched in the last decade. The next barplot shows the distribution of the satellites to countries in the last decade (not including satellites which are still operational):

We decided to focus on United States, Russia and China, for they are leading in the amount of satellites, in a noticeable gap form the other countries. Next, we examined the distribution of the days in space of the three countries:

One can observe that the distribuitions of the days is space are not Normal. Afted an application of concave transformations on the data, the distibuition looks like:

Now part of the ditribution seems like taken from a GMM (gaussina mixure model) with 2 gaussians[3]. Neverthelss, this is still not normal, so the classic t-test and F-test won't bring any help here.

[1] https://www.kaggle.com/datasets/heyrobin/satellite-data-19572022

[2] As reported in Kaggle, the dataset gets updated mounthly.

[3] This might imply that the population of the satellites can naturally divided to 2 sub-populations. For example, satellite that manufactured with flaws that the centroid of their disributed around small number of days, compared to the group of satellites that properly manufactured.

Methods

First, we will examin

The test statistic

Results

Discussion

Refrences

Appendix: The code