Deep dive into ANOVA(Analysis of Variance) Test

Van Nguyen
Department of Computer Science
Boise State University
vannguyen599@u.boisestate.edu

1 Introduction

Kickstarter is an online platform that enables creators to raise funds for their innovative projects from a community of supporters. The platform has become significantly popular among other crowdfunding platforms due to their user-friendly interface and dynamic ecosystem that fosters diverse creative campaigns. As someone who has backed several Kickstarter campaigns and found the experience satisfying, I at one point have considered launching some experimental project related to keyboards on the platform. Like other creators, the ultimate question is how to ensure their projects to be funded successfully. Given this question, we need to engage in a study to understand what factors can influence the success of a project. The objective for this course project is to analyze all Kickstarter project data to identify patterns in order to predict successful funding campaign. The motivation is to employ data science techniques to analyze and uncover actionable insights that can inform future crowdfunding strategies.

2 Methodology

The datasets used will be generated by *https://webrobots.io/kickstarter-datasets/* which a platform where they have a regular scraper robot that crawls all Kickstarter projects every once a month. I used the datasets through 2025-10-13, which have 186,372 data points and 46 columns

2.1 Data Exploration

Kickstarter does not provide any public API to retrieve data other then *https://status.kickstarter.com/api/v2/summary.json* which is used to look up the status of the platform. As a result, it is necessary to explore and investigate existing datasets from other platforms such as **Kaggle** to determine whether they are relevant or suitable to use. The selected dataset should be up-to-date and include all key factors required for an examination of funding trends and project success. I have attempted to scrape the data directly from the website but was not able to get all the sufficient data so I have decided to go with *https://webrobots.io/kickstarter-datasets/*. However to make sure that data is relevant and webrobots does not just patch missing data with wrong information, I have matched data I scraped with what currently provided by webrobots on 5-6 projects. **sample.json** is the data that I have pulled down to check for webrobots's validity.

Other than that we need to go through the dataset in order to make sure that we only select our features for data training. After consideration, we decided to go with the below list as features:

- The campaign's financial goal in USD
- The time taken to prepare the campaign (from the creation date to the launch date)
- The duration of the campaign (from the launch date to the deadline)
- The geographical location of the campaign (whether it is based in the US or not)
- The category under which the campaign falls

We will need to do some data preprocessing where we dropped unused columns and check for duplicates or invalidate data points. Other than that for any timestamp we want to convert them to datetime object. In order to accommodate the above feature list, we need to create extra columns that suit our purpose namely *main_category*, *sub_category*, *campaign_goal_USD*, *campaign_location_US*, *campaign_timeline*, *prep_timeline*.

2.2 Data Analysis

We will use the sample data to compare different class type to evaluate how they affect the student

exam scores using one-way ANOVA. All of the steps used in this paper used the guidance from (Thevapalan, 2024)

1. Define the hypothesis which are the Null Hypothesis (H_0) and Alternative Hypothesis (H_1) where

$$H_0 : \mu_{\text{hybrid}} = \mu_{\text{in-person}} = \mu_{\text{remote}}$$
$$H_1 : \text{At least one } \mu \text{ is different}$$

where $\mu_{i...3}$ represent the means of different class type

2. Before getting into the calculation, we need to have these assumptions straightforward:

- The student's score in one class type should be independent of the other class type
- The variance within each group is equal (we generated 100 data points for each of the class type category)
- The data within each group follow a normal distribution

If any of the above assumptions violated, the final result maybe invalid.

3. Next step, we will get all the calculation needed for ANOVA Below is the mean for each of the class type (A_i)

$$A_i = \frac{\sum_{\text{Group}} x}{n_{\text{Group}}}$$

$$A_{\text{hybrid}} = 52.96$$

$$A_{\text{in-person}} = 53.54$$

$$A_{\text{remote}} = 50.11$$

Then, we get the overall mean (G)

$$G = \frac{\sum x}{N}$$

$$G = 52.2$$

Finally, we calculate the sum of squares for each group

$$SS_{Group} = \sum_{Group} (x - G)^2$$
$$= \sum_{Group} x^2 - \frac{\left(\sum_{Group} x\right)^2}{n_{Group}}$$

$$SS_{hybrid} = 368700 - \frac{5296^2}{100} = 88223.84$$

$$SS_{in-person} = 379702 - \frac{5354^2}{100} = 93048.84$$

$$SS_{remote} = 332937 - \frac{5011^2}{100} = 81835.79$$

4. After the training, we will fill out the values in the below table so we can refer to it later

Model type	accuracy	precision	recall	f1	roc
Logistic	0.690141	0.701068	0.866327	0.774985	0.712753
Decision Tree	0.643219	0.717607	0.693758	0.705481	0.631295
Random Forest	0.695077	0.734781	0.790148	0.761459	0.730677

Table 1: Summary of chosen model's performance before tuning

$$SS_{\text{between}} = \sum_{\text{Groups}} n_i (A_i - G)^2$$

$$SS_{\text{between}} = 674.07$$

$$SS_w = \sum_{\text{Groups}} SS_i$$

$$SS_w = 263108.47$$

We then can use SS_{total} to verify if our calculation is correct or not by adding SS_{between} and SS_w to see if they are the same as the calculated SS_{total}

$$SS_{\text{total}} = \left(\sum x^2\right)_{\text{total}} - \frac{((\sum x)_{\text{total}})^2}{N}$$
$$SS_{\text{total}} = 263782.596$$

Next we calculate the mean square within class types and between class types

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}}$$

$$MS_{\text{between}} = \frac{674.07}{2} = 337$$

$$MS_w = \frac{SS_w}{df_w}$$

$$MS_w = \frac{263108.47}{297} = 885.88$$

df_{between}: the number of class types minus one df_w: the number of class types subtract the total number of participants Lastly, we calculate the F-statistic and p-value

$$F = \frac{MS_{\text{between}}}{MS_w}$$
$$F = \frac{337}{885.88} = 0.3804$$

We can see that the number we got the same as the F-statistic value we got from `scipy.stats.f_oneway` (this can be found from the attached notebook), which mean the calculated p-value is going to be 0.684

5. The found F-statistic is quite small, normally a higher F-statistic represents a greater disparity between group means compared to randomness. The found p-value is greater than the predefined threshold(typically 0.05) which means that our null hypothesis is correct and we can conclude that at least one class type has more impact the exam scores.

2.3 Post-hoc test

Even though, ANOVA tells you whether or not there is a disparity between group means, but it does not indicate which specific groups differ significantly. This is where post-hoc tests come in—they conduct pairwise comparisons to pinpoint the exact groups with significant differences. These tests are crucial when dealing with more than two groups and the ANOVA result is significant. For instance, after performing a one-way ANOVA on students' score across three class type (Hybrid, In-person, and Remote), the test identifies a somewhat difference but no other information about group differs significantly. With a post-hoc test, we can determine which class types differ in their impact on exam scores. Commonly used post-hoc tests include Tukey's Honestly Significant Difference (HSD) (Howell, 2010) and the Bonferroni Correction (Armstrong, 2014).

3 Conclusion

In this paper, we explored the fundamental concepts and applications of the Analysis of Variance (ANOVA) test especially one-way ANOVA. ANOVA is a powerful statistical tool used to determine whether there are significant differences between the means of multiple groups. We discussed the core concepts, including getting all the assumptions required for valid results, and interpreted the final result(including whether or not the null hypothesis is accepted). Additionally, we emphasized the importance of post-hoc tests, such as Tukey's HSD and the Bonferroni Correction, in identifying specific group differences when the ANOVA results are significant. Overall, ANOVA remains an essential technique in experimental design and data analysis, providing researchers with a robust method for comparing group means and uncovering insights in diverse fields such as education, agriculture, and healthcare. By understanding its principles and limitations, practitioners can effectively apply ANOVA to real-world problems and make data-driven decisions.

References

Robert A Armstrong. 2014. When to use the bonferroni correction. *Ophthalmic and Physiological Optics*, 34(5):502–508.

D. C. Howell. 2010. *Statistical Methods for Psychology*, 7th edition. Wadsworth, Cengage Learning.

Arunn Thevapalan. 2024. Anova test: Definition, types, examples, and assumptions. Accessed: 2025-10-27.