Data Analysis of Wellspring Member Engagement

You can either put a subtitle here or delete this line

, TUT0104, and 0104-D

Introduction

- The objective for this project is to
- We are provided with the “ ” dataset
- Our target population for presenting the results is the

Objectives

List of research questions

For this investigation we have chosen three research questions,

- Is Wellspring’s membership age distribution balanced?
-
-

Question 1: Introduction

- Research question - Is Wellspring’s membership age distribution balanced?

Question 1: Data Visualization

- Elder Members (40) are the majority, comprising 52.1% of the total membership.
- Young Members ($<$40) are significantly underrepresented, making up only 6.9% of members.
- A large portion of members fall into the elder category, suggesting that Wellspring's services are more utilized by older individuals.
- 41% of members have missing age data (`NA`). This means almost half of the data set lacks birth year or age information. The high percentage of missing values could affect conclusions about age distribution.

Question 1: Statistical Analysis

- H (Null Hypothesis): The proportion of young members is equal to 50% (i.e., the age distribution is balanced).
- Ha (Alternative Hypothesis): The proportion of young members is not equal to 50% (i.e., there is an imbalance in the age distribution).
- The test statistic and the p-value can be estimated from a graph based on multiple repetitions of simulated samples assuming the null hypothesis is true.
- The test included **10,000 simulations** to determine the distribution of the proportion of young members under the assumption that the true proportion is 0.5.

Question 1: Results

- The **p-value** was found to be **0**.
- Since the p-value is **0**, which is less than 0.001, we conclude that we have **very strong evidence against the null hypothesis** that the proportion of young and elder members is 50% each.
- The age distribution of Wellspring members is significantly **imbalanced**, with far fewer young members than expected under a balanced 50/50 assumption.
- Further research is suggested to investigate possible causes behind the underepresentation of younger individuals at Wellspring, and whether targeted outreach or program redesign could help attract a more age-diverse membership.
- However, if the missing data is random, then we can still be reasonably confident that the analysis represents the overall membership. Otherwise it can introduce serious bias into the analysis. (young people are reluctant to disclose their age (more likely to be missing). Then the analysis then

Question 2: Introduction

- Research question - **Is the proportion of the first listed property a travel ad, equal to 50%?**
-

Question 2: Data Visualization

The above bar graph shows the frequency of the first listing being a Travel Ad or not. It can be observed that out of the sample data, 600 of the first listings were not Travel Ads, while the rest 400 were Travel Ads.

Question 2: Statistical Analysis

- Null Hypothesis (H_0): Among all the searches on the Expedia website that span the period from 2021-06-01 to 2021-07-31, the proportion of the first listings which are advertised in travel ads is equal to 50%.
-

Question 2: Results

- The actual proportion of first listings advertised in a travel ad was found to be 0.419.
-

Question 3: Introduction

- The variables member_start_year, member_start_month were used to determine the registration date for each member.
- Taking the minimum of the variables delivery_year, delivery_month, delivery_day, as well as filtering by Present in attendance_status were used to determine the date of the member's first attended event.
- The new variable system_change was created to differentiate between if a member registered before or after the system change date, and grouped accordingly.

Question 3: Data Visualizations

The above bar graph shows the proportion of members attending their first event within 30 days. Before the registration system change, 67.8% of members attended their first event within 30 days, while only 35.5% did after the change.

Question 3: Statistical Analysis

- Null hypothesis (H_0): The proportion of members attending their first event within 30 days of registration is the same before and after the registration system change.
- Alternative hypothesis (H_1): The proportion of members attending their first event within 30 days changed after the registration system change.
- To assess the significance of the observed difference in proportions between the two groups, we performed a permutation test with 10,000 random permutations. This included randomly shuffling the system_change labels to simulate the null hypothesis, and recalculating the difference in proportions for each permutation to build the null distribution of differences.

Question 3: Results

- The actual difference between the proportions was -0.32564096245674. This indicates a 32.6 percentage point decrease in participation within the first 30 days after the system change.

  The p-value was found to be 0.

  Since the p-value is less than 0.05, we conclude that we have very strong evidence against the null hypothesis.

  We conclude that the system change had a statistically significant negative effect on member participation.

Conclusion

- From the bootstrapping investigation we can state with 95% confidence that the mean stay length for listings on the Expedia website in the specified timeframe is between 2.897 and 3.258 nights.
    - In comparison to Expedia's competitor, Airbnb, it is a lower average stay length with Airbnb averaging 3.9 nights per customer, hence we have decided to analyze certain factors which may influence this.
- From the proportion based hypothesis testing we can conclude that we have very strong evidence against the fact that among all the searches on the Expedia website in the specified timeframe, the proportion of the first listings which are advertised in travel ads is equal to 50%.
- We can conclude that the option of first listings having free cancellation does lead to a difference in the average review rating according to the two proportion hypothesis which showed that there is strong evidence against there being no difference between the average review rating for first listings between the groups which have free cancellation and don't have free cancellation

Limitations

- For this investigation specifically questions 1 and 2, only thefirst listings we used so we can't necessarily generalize the findings for the whole dataset hence further analysis is needed.
- The data set was only for a small duration of time specifically during the covid-19 pandemic hence travel restrictions to countries would have had heavy influence on the data collected.
- The possible factors in the data dictionary which can influence listings chosen such as free wifi and breakfast were removed in the data set given hence possible other reasons for the indicator variable specifically in question 3 could not be recorded

References and Acknowledgements

- Airbnb Economic Impact. Retrieved March 20th, 2022, from https://blog.atairbnb.com/economic-impact-airbnb/#:~:text=Airbnb%20guests%20stay%20on%20average,%24713%20for%20the%20average%20visit

- The authors will like to thank Colin for their helpful suggestions on the visualizations used for the 2 proportion hypothesis tests and simplfying our bootstrapping interval approach for this project
- The authors will like to thank Uzair for their helpful suggestions in reevaluating our research questions including solidifying our approach for the porportion based hypothesis testing and overall selection of the 3 research questions