The Greatest Suggestions ever for RCYC

LingXin Li, tutorial group(my ta:Gloria)

April 1, 2021

Introduction

The data I used is a random sample of 1,000 Royal Canadian Yacht Club (RCYC) members, and it includes information for each of them in year 2017 and year 2020.

- My goal is to help Royal Canadian Yacht Club (RCYC) further explore the conditions of their services that they have provided for their members. After all, they can have a deeper understanding of the patterns, so that they can make improvements and development in order to serve their members better.
- audience: boss/managers of RCYC company

Objectives

Here is the research questions that I want to investigate for the purpose of helping RCYC to make improvements for company and consumers.

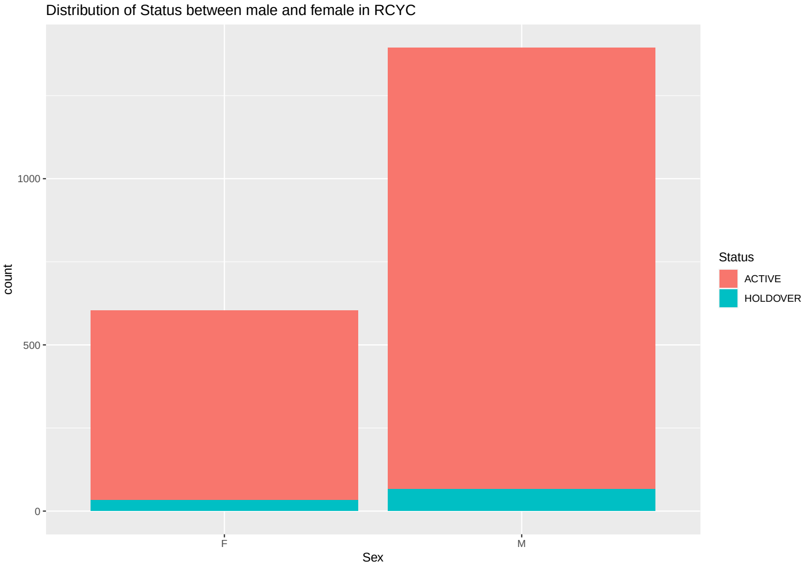
Research Questions	Population
1.proportion of “Holdover” between male and female RCYC members.	All RCYC members at all time.
2.average total spending per member.	All active members of RCYC in 2017.
3.relationship between city_dining and other_spending.	All active members of RCYC in 2017.

Data Summary

I cleaned the data by:

- creating a tibble called “Active_2017” which only includes the active RCYC members in 2017 in the sample data.
- creating a variable called “total_spending” which is a sum of city_dining, island_dining, bar_spending and other_spending per active RCYC member in 2017.
- (people who have 1 or more missing values from the following:“Sex, city_dining, island_dining, bar_spending, other_spending” are removed for above)

Data Visualization for research q1



Statistical Methods for research q1

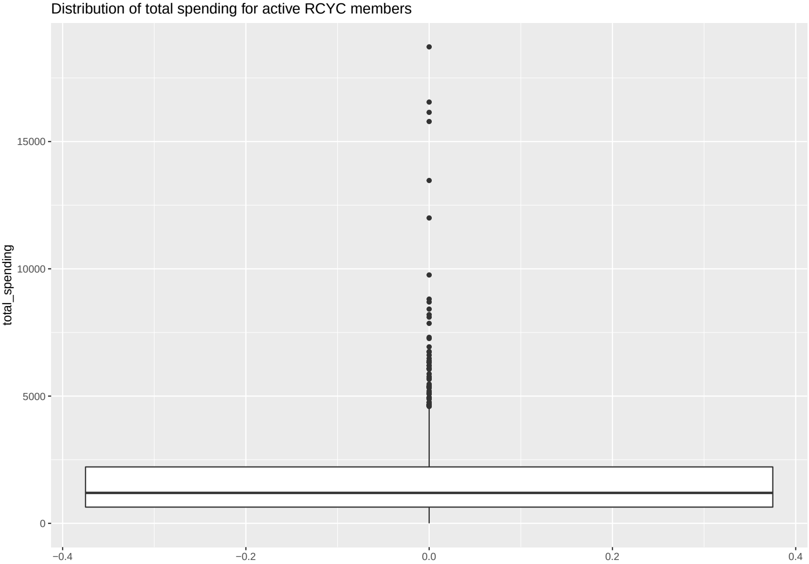
I will use two sample hypothesis test (Randomization test) to compare the proportions of “holdover” members between proportion of “holdover” of male in RCYC and proportion of “holdover” of female in RCYC at all time.

- $H_0 : p_{male} = p_{female}$, $H_1 : p_{male} \neq p_{female}$
- p_{male} is the proportion of male in RCYC who put their membership on hold for that year in all male members in RCYC at all time. p_{female} is the proportion of female in RCYC who put their membership on hold for that year in all female members in RCYC at all time.
- H_0 is the null hypothesis assuming that the proportion of status “HOLDOVER” for members in RCYC of all of the years is the same between female and male at all time. H_1 is the alternative hypothesis that proportion of status “HOLDOVER” for members in RCYC of all of the years is not same between female and male at all time.

Results for research q1

- I found that the sample difference(test statistics, which is what we get when we do an experiment in real world) between proportion of male RCYC members whose status is "HOLDOVER" at all years in the sample and proportion of female RCYC members whose status is "HOLDOVER" at all years in the sample is -0.00729.
- the corresponding p-value 0.485 shows we have no evidence against the null hypothesis assuming that the proportion of status "HOLDOVER" for members in RCYC of all time is the same between female and male at all time. Type 2 error may occur which we might failed to reject H_0.

Data Visualization for research q2



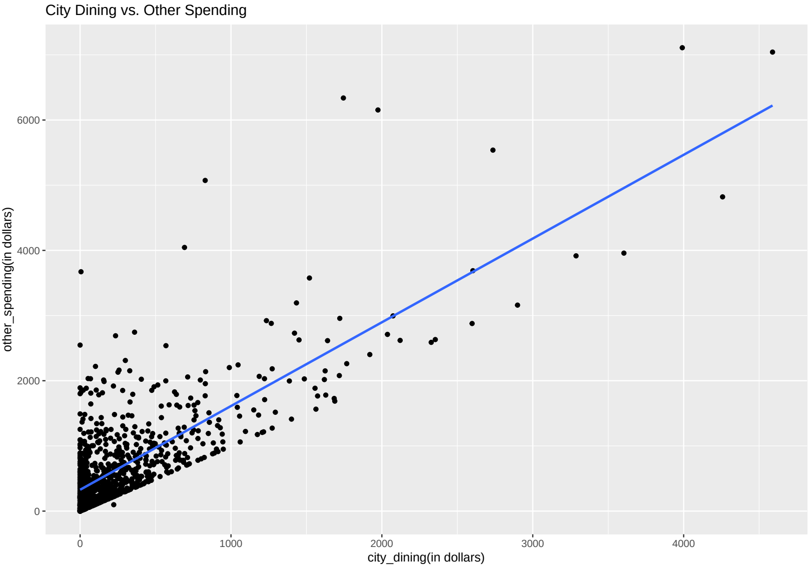
Statistical Methods for research q2

- Interested in the true population parameter: true mean of total spending per active member in 2017
- I will use method of percentile bootstrap method, in order to get a confidence interval, which can create a range of plausible values for the true population parameter.
- I will calculate a 97% confidence interval for the mean total spending in dollars for members.

Results for research q2

- I found that we are 97% confident that the true mean total spending per active member in 2017 is between 1610.073 dollars and 1872.116 dollars. In other word, I expect 97% of the interval[1610.073, 1872.116] in dollars contains the true mean total spending per active member in 2017.

Data Visualization for research q3



Statistical Methods for research q3

- I will make a simple linear regression model to explore whether or not city dining(2017 yearly amount spent on dining at the RCYC's restaurants in the city of Toronto) is a useful predictor to predict other spending.
- In other words, I want explore whether or not there is a linear association between city dining and other spending.

$$y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i$$

$$other_{spending} = \beta_0 + \beta_1 \times city_{dining} + \epsilon_i$$

- $H_0 : \beta_1 = 0$, and H_0 is the null hypothesis assuming that there is no linear association between other spending and city dining.
- H_1 is the alternative hypothesis that tells there is an linear association between other spending and city dining.
- β_1 is the average change in other spending for 1-unit change in city dining.
- β_0 represent the amount of dollars of other spending when city dining is 0 dollar. ϵ_i is the random error term for every observation since we do not know the true values of β_0 and β_1.

Results for research q3

The estimated simple regression of other spending on city dining is:
$other\hat{}_{spending} = \hat{\beta}_0 + \hat{\beta}_1 \times city_{dining}$

$other\hat{}_{spending} = 327.44510 + 1.28462 \times city_{dining}$

- The estimation of β_0 is $\hat{\beta}_0$ which is 327.44510. This tells us when city dining is 0 dollar, the average other spending on is 327.44510 dollars.
- The estimation of β_1 is $\hat{\beta}_1$ which is 1.28421. This tells us each additional dollar in city dining is associated with an increase of 1.28421 dollar in mean other spending based on the association we observe.

- p-value for the null hypothesis assuming that there is no linear association between other spending and city dining is $(4.7902043 \times 10^{-191})$ which is extremly small. Based on the p-value, I conclude that we have strong evidence against the null hypothesis.
- Therefore, city dining is an useful predictor to predict other spending, and there is a linear association between other spending and city dining. (To be more details, a moderate positive linear association can be observed according to the data visualization without calculating p-value.)

Limitations

- q1 include data in year2017 and year2020, where year2020 involved here is a special rare pandemic year, decisions for people might be special and changed for this year, so that the results might be relatively different as usual.
- sample size of 1000 not a big sample size, the result might be relatively less accurate than a bigger sample size or compared to the result for the true population parameter that we would never know.

Conclusion

Based on the result for:

- q1: I suggest when RCYC want to make improvement or decisions, they could consider the services for female and male equally and improve both at the same time, because both gender are likely to have the same chance to put the membership on hold.
- q2: It depends on RCYC managers or boss to decide whether or not the range is within their expectations, and to keep the services city&island&bar&other services, according to whether or not they think that plausible range for the true mean total spending is an appropraite and reasonable range.
- q3: I suggest to also focus on city dining if you want to increase other spending, because this is relatively a much more efficient way to consider since city dining is a useful predictor to predict or make inferences about other spending and there is a linear association between them.

References and Acknowledgements

The authors would like to thank Prof Moon, Prof Caetano, ta Gloria, ta Yaru, ta Amar for helpful suggestions and comments that improved the presentation of this poster.