Demographic analysis for Royal Canadian Yacht Club

Qianning Lian, Naihe Xiao, Minglang Jin, Zixiang Zhang group 64

April 1, 2021

Overall Introduction

To gain highest economic profit and allocate resources more efficiently, a company, especially a recreational company, must have comprehensive knowledge about their members.
In order to help Royal Canadian Yacht Club(RCYC) better understand the demographic structure of their customers, we raised three research questions to make inferences about the RCYC members, each concerning a different aspect(sex, consumption, age) of the population, and thus yield results that could provide meaningful information for RCYC managers. This project can not only help the managers acquire higher revenues, establish a more efficient, secured, and reasonable system but also improve the experience(and very possibly healthiness) of the consumers.

Data summary

The population(collect of people we are interested in)we made inference about is all RCYC members, using a sample of them from 2017 to 2020, and our audiences are managers of the Royal Canadian Yacht Club. To approach our population and parameter(feature of a popoulation we are interested in)for question 1, we created a new tibble, RCYC1, including only members with "ACTIVE" status in the sample and focused on variables concerning sex and active status. For the research question2, we considered variables about consumption in Toronto city restaurants and participation in fitness activities. Since there are none values in both columns, we filtered them out and stored the rest of the data in dataset RCYC2. To investigate the third question, we only kept members who used fitness facilities(thus has value "Y" in "fitness" variable) and saved them in tibble called RCYC3.

Below is the first 6 rows(members) of each new tibble we created, with table1 referring to RCYC1, table2 referring to RCYC2, and table3 referring to RCYC3.

Table1		Table2		Table3	
Sex	Status	fitness	city_dining	fitness	
M	ACTIVE	N	137	Y	72
M	ACTIVE	Y	829	Y	46
M	ACTIVE	N	0	Y	59
M	ACTIVE	Y	660	Y	62
M	ACTIVE	Y	403	Y	33
M	ACTIVE	Y	559	Y	48

Research question 1

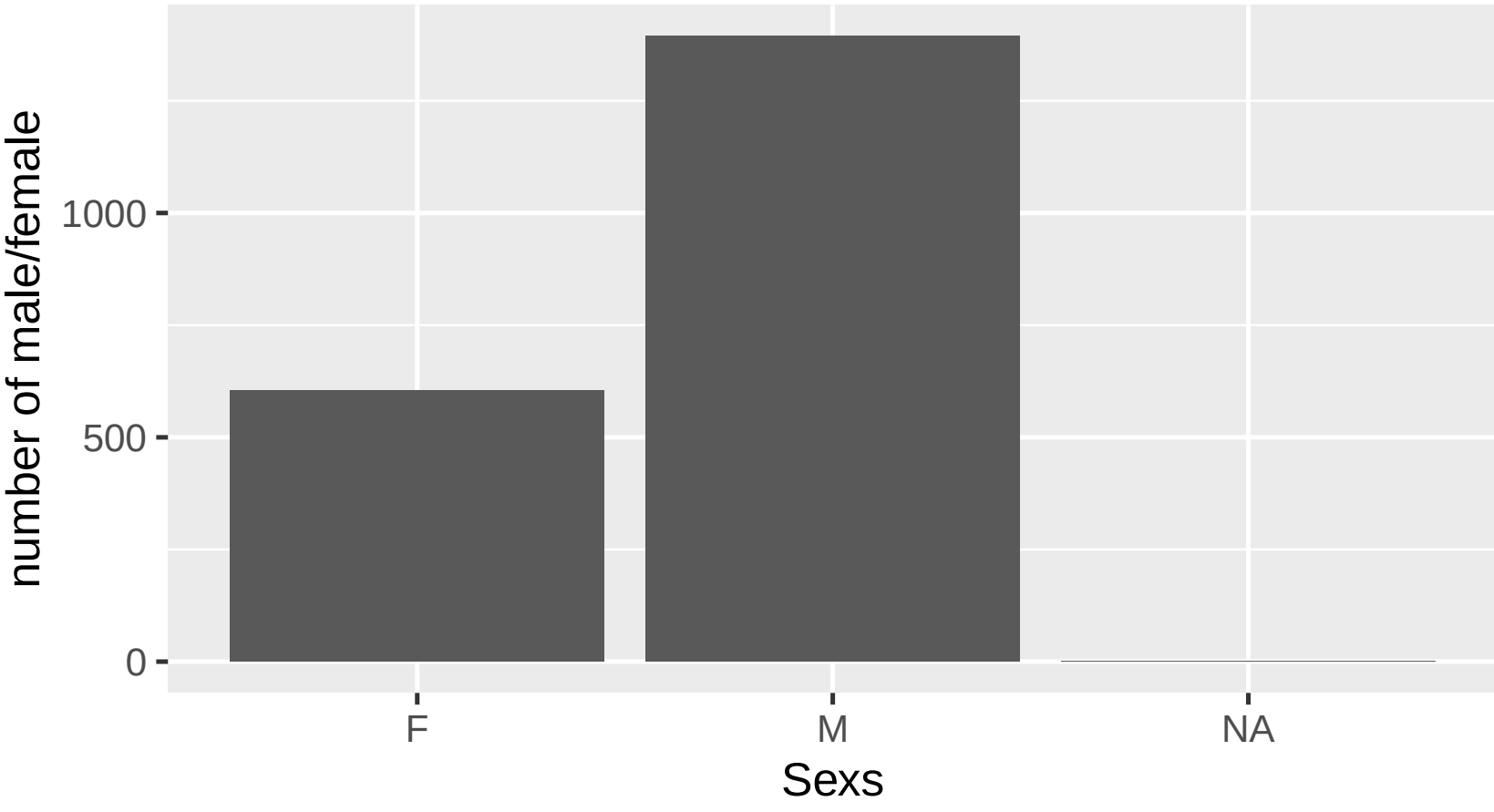
Our first research question is sex-related: **"Is the proportion of males in members with "ACTIVE" status (members active in a particular year) in the RCYC 50%?"** This question makes inferences about all members of the RCYC with "ACTIVE" status from 2017 to 2020.

We came up with this question after noticing the huge difference between the number of female and male members in RCYC(**604** and **1394**, respectively) in the sample data, and wondered if the same proportion shows in members with "ACTIVE" status, which is a more direct and meaningful indicator of members' usage of facilities. We assumed that the real number of male members with "ACTIVE" status would be equal to that of female members.

This question provides valuable information for RCYC managers, who would be capable of better allocating their advertising resources and facility concentrations for different genders.

Since the sex variable is **categorical**, we took it as the x-axis to generate a barplot to visualize the distribution of the sample data, as shown in Figure1. We can see from the distribution that the number of male members with "ACTIVE" status is approximately 2 times of the female members.

Figure1



We applied **hypothesis test for one proportion** since we are testing a specific parameter of a population. We first decided our null and alternative hypothesis:

$$H_0 : p = 0.5$$

$$H_A : p \neq 0.5$$

Where p refers to true proportion of male in "ACTIVE" RCYC members from 2017 to 2020. We selected a **0.05 significant level**(the value to evaluate rejection of H_0), and attained the test statistic(help decide compatibility with H_0) from the wrangled dataset RCYC1.

n	test_stat
1899	0.6993154

Assuming H_0 is true, we simulated 1000 samples of size 1899 and calculated statistic for each sample, collecting them with an object.

Finally we evaluated the evidence against our null hypothesis by calculating the probability of observing data that are at least as extreme as the sample data in the set of statistics of our 1000 samples(also called the p-value).

p_value
0

The p-value happened to be 0, meaning that there are no cases more extreme than our test statistic in the 1000 samples. Based on this p-value, we concluded that we have **strong evidence** against our null hypothesis, and since we picked 0.05 as our significance level, we rejected our null hypothesis, which is the true proportion of male in "ACTIVE" members in RCYC from 2017 to 2020 is equal to 50%.

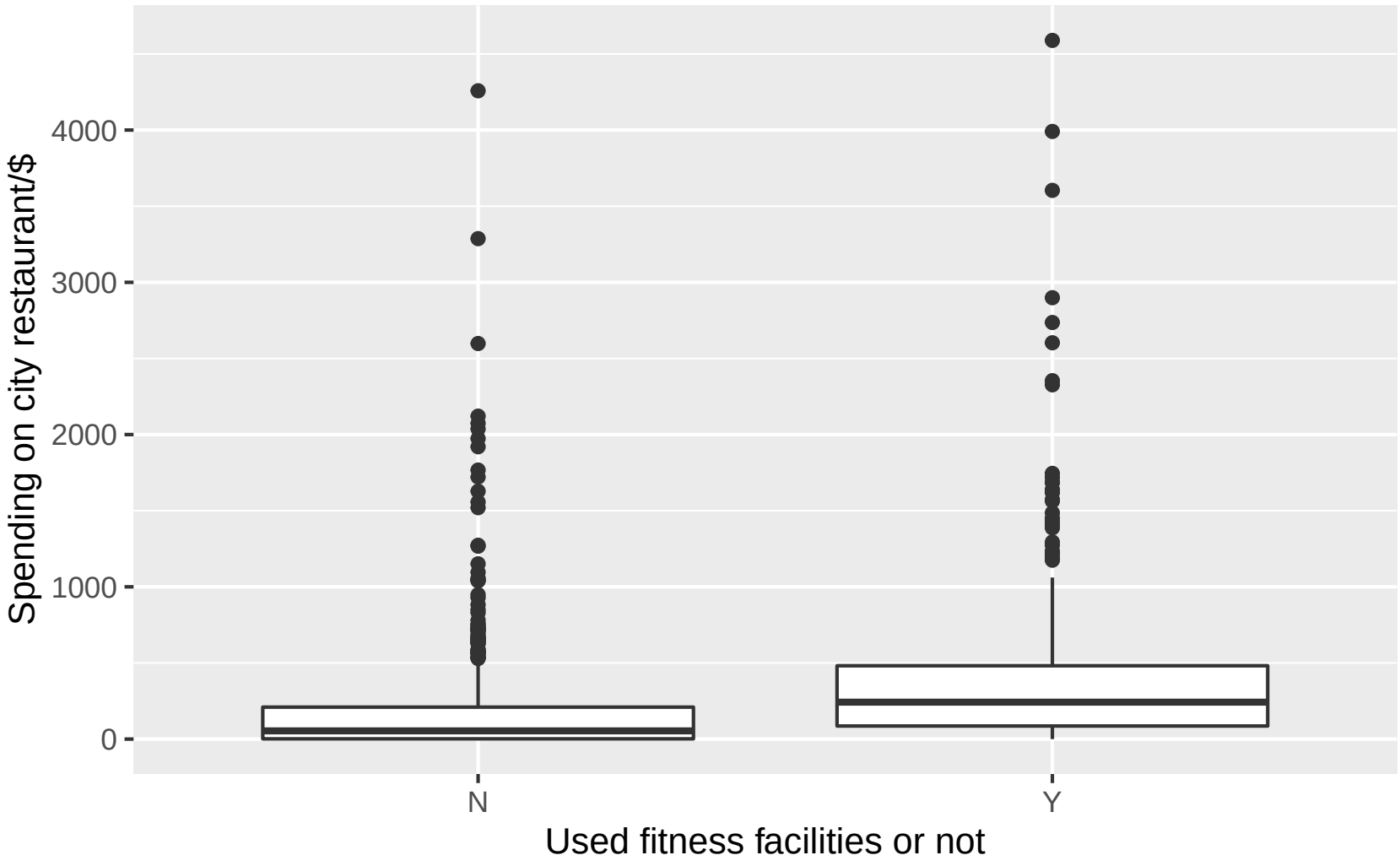
Research Question 2

Our second question is: **Is the mean yearly consumption on RCYC restaurants in Toronto city by members who use its fitness facilities equal to the that of members who don't use the fitness facilities?**

Our inspiration for this question arose when we realized the necessity for managers of RCYC to pay attention to the quality and accessibility of their city restaurants. Since fitness is one of the most essential activities in RCYC, the managers had better know its relationship with restaurant consumption to improve the restaurant's possible deficiencies.

Since the fitness variable is **categorical** and city_dining is **numerical**, we applied a side-by-side boxplot, as shown in Figure2, from which we noticed that the mem-
ber who use fitness facilities tend to spend more than those who don't.

Figure2



We used the **randomization test** because we are testing a parameter of 2 categories, and selected the **significant level to be 0.05**. Then we established our null hypothesis and alternative hypothesis.

Null hypothesis (H_0): The members who use fitness facilities have the same average consumption in RCYC's city restaurants to the members who do not use fitness facilities.

$$H_0 : \mu_{use} = \mu_{notuse}$$

Alternative hypothesis (H_1): The members who use fitness facilities do not have the same average consumption in RCYC's city restaurants to the members who do not use fitness facilities.

$$H_1 : \mu_{use} \neq \mu_{notuse}$$

We then attained the test statistic(sample parameter), and conducted 1000 simulations, from which we collected the difference in city_dining between the two groups of people.

test_stat
227.8873

After the simulation we calculated p-value, the probability of observing data that are at least as extreme as our test statistic, to evaluate the extremeness of our test statistic(227.8873).

p_value
0

We noticed that the p-value is very close to 0, implying that none of our simulations under the null hypothesis were as extreme as our test statistic. Thus, we have ***very strong evidence*** against our null hypothesis, and since we picked 0.05 as our significance level, we rejected our null hypothesis that there is no difference between the mean spending on dining for members who use the fitness facilities and members who do not use.

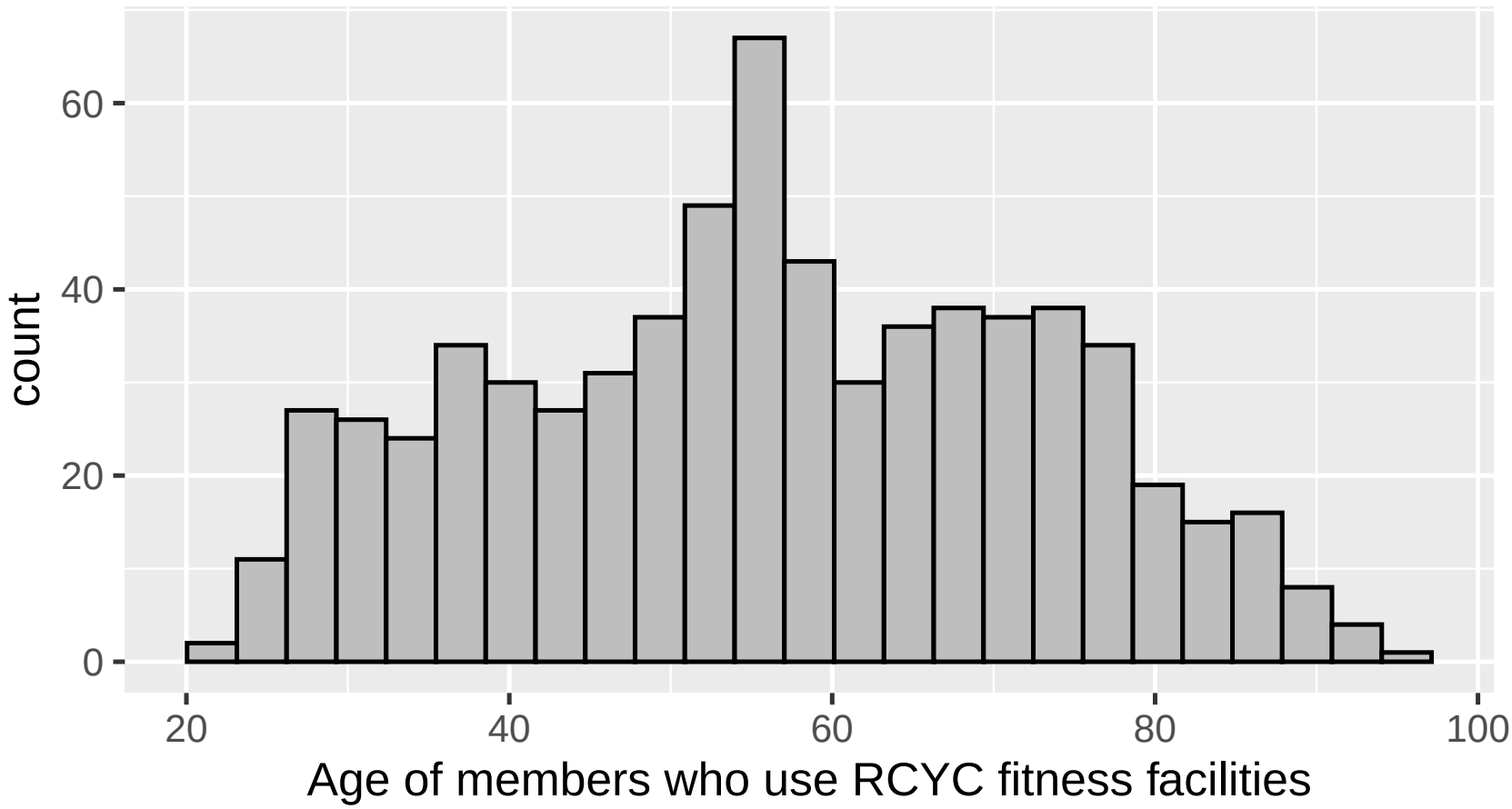
This result is valuable as it reveals the fact that members using the fitness facilities are very likely going to spend more on city dining than those who have no interest in fitness. Therefore RCYC managers should probably renovate their city restaurants to make them more suitable and accessible for fitness enthusiasts, for instance adjusting the menu and building more locker rooms inside the restaurants.

Research Question 3

Our third question is: **"What is a range of plausible values for the mean age of RCYC members who use its fitness facilities?"** This question makes inferences about all RCYC members who use RCYC fitness facilities, and it helps the managers determine the style of these fitness facilities according to the interval of major clients' ages. We maintained our concentration on fitness facilities to provide a thorough analysis on this particular group of people, which enable us to provide deeper and more deliberate suggestions.

We applied a histogram to visualize the distribution of member ages since they are **numerical** using our wrangled dataset RCYC3, as shown in Figure 3. The distribution is generally symmetrical with no obvious skewness, has one mode around 53, and values are evenly spread out on both sides.

Figure 3



We used the **bootstrapping method** as we want to generate an interval for the parameter we were predicting. Based on the original sample which contains 684 members(assuming it's a representative of all members using fitness facilities), we estimated a distribution of our parameter by re-sampling 2000 samples with replacement(meaning that same member from the original sample can occur more than once in the new sample). We attained statistics of each sample, and entered the next phase which generated a range of possible values for our true parameter.

We chose **95%** as the confidence level as it included most values, from which we computed the corresponding confidence intervals: [55.12262,57.54971].

Table 6: Table3

	x
2.5%	55.12262
97.5%	57.54971

This means that if we start with many different samples of 684 members using fitness facilities and getting a bootstrap confidence interval based on each one, 95% of these confidence intervals would include the true mean value. A more formal conclusion is that we are 95% confident that the true mean age of all RCYC members who use fitness facilities is between 55.12262 to 57.54971. Based on this value, we suggest RCYC managers build fitness facilities safer and more convenient to ensure security and attract more customers of these ages.

Limitations

Although we have drawn rational results from all three questions, there are, more or less, restrictions on each conclusion. We rejected our null hypothesis based on the 0.05 significance level in both questions 1 and 2. However, it's possible for the Type1 error to occur, in which case we observed a very unusual outcome.

We assumed the sample to be a good representative of the entire population in question3, which might not necessarily be. If it's indeed a biased sample, our confidence interval would also be biased, and our prediction would be inaccurate.

Overall conclusion

Despite the above limitations, we had significant findings of the demographic features of RCYC members. We concluded that male members tend to participate in club activities more actively. People interested in fitness-related events are likely to consume a more significant amount in Toronto city restaurants. In addition, we estimated that the ages of people who use fitness facilities are approximately centered between 55 and 58. In general, we can conclude that members aged around 55 to 58 are more inclined to enroll in the fitness center and spend money on dinings, and male members will possibly outnumber female members.

Consequently, we suggest that the RCYC managers design more suitable and attractive activities for women to enlarge their market and potential customers. For example, yoga classes in its fitness center and aquatic amenities near its dock. The managers can also open restaurants of different styles around its fitness center so that members have more choice and might bring their family members for dinner.