THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF STATISTICS
ANALYSIS OF STATISTICAL CONSULTING DATA BY USING TEXT MINING

YIYANG WANG
SPRING 2020

A thesis
submitted in partial fulfillment
of the requirements
for baccalaureate degrees
in Statistics and Mathematics
with honors in Statistics

Reviewed and approved* by the following:

Le Bao
Associate Professor
Thesis Supervisor

David Hunter
Professor
Director of Online Programs
Honors Adviser

*Signatures are on file in the Schreyer Honors College.

Abstract

Working in the information age, people are embracing an overabundance of data. In order to properly analyze the research data and make correct decisions, people use statistics more and more extensively, and often need help with statistical analysis to gain a better understanding of their data or address the scientific questions they have. The Statistical Consulting Center (SCC) is a professional place where people can get help. The clients coming to academic statistical consulting centers usually provide a brief description of personal background, such as names, emails, departments they belong to, and the research questions before the meeting. These information help the consultant better prepare the consulting meeting. After the meetings, the consultant writes the meeting summary, which includes the suggestions and recommendations. This thesis is mainly focusing on analyzing the relationships between variables in the client information and the consultant summary.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my honor thesis supervisor, Dr. Le Bao, for his guidance and support during the process of writing this thesis. I would also like to thank Dr. Maggie Niu for providing the data and resources.

I would like to thank my honor advisor, Dr. David Hunter, for his consultation and guidance.

I would like to thank Schreyer Honors College for providing me this opportunity.

I would like to thank all the statistics and mathematics professors for teaching me the knowledge I have ever learned and directing me to find the career I would like to pursue in my future.

I would like to thank my families and friends for their unconditional support and love.

Chapter 1

Introduction

1.1 What is Statistical Consulting Center?

Statistical Consulting Center (SCC) is an organization providing professional suggestions and assistance in statistical design and analysis. In the era of big data, analyzing data using statistical methods is important for various industries and hence, Statistical Consulting Center becomes more and more important and popular. SCC helps clients in diverse fields and areas, such as medicine, law, business and sociology. There are around 100 university statistical consulting centers in the United States of America now[1].

1.2 Statistical Consulting Center at Penn State University

Statistical Consulting Center at Penn State University is operated by the Department of Statistics and it aims to provide statistical support to all university students, faculty, staff and external industry staff and government officials. Graduate students majoring in statistics serve as consultants and handle real life statistical research cases. The particular areas that the Penn State Statistics Consulting Center provides support in are survey sampling, statistical software programming, data analysis and result interpretation[2]. The statistics consulting center provides support in two forms:

free short-term consulting and long-term collaboration, which depends on the amount of help that the client needs. Before the consulting meeting, the client needs to complete the online "Consulting Request Form", which asks for client's basic information, the statistical question description and client's research information. After submitting the request form, the client can pick an available meeting time slot. The background information provided by the client is then reviewed by the consultant to better prepare the meeting. After the meeting, the consultant writes a summary and records the meeting details.

1.3 Motivation

Although the statistical consulting center helps with many researches in multiple areas and departments, there are few projects that focus on analyzing data collected in statistical consulting center. As the importance of SCC is in sustainable growth, knowing more about SCC and improving its efficiency become necessary. Sufficient information benefits both clients and consultants. On one hand, it gives the consultant a direction to prepare better for next clients. For example, understanding the distribution of clients status helps the consultant to locate the statistical knowledge level of the next client; knowing about the distribution of the departments clients belong to helps the consultant to do specific readings to get familiar with the topics they will talk about during meetings; learning about the distribution of consultation keywords helps the consultant to prepare some topic-related resources for the client's further interest. On the other hand, this information also helps the client. For instance, knowing the distribution of follow-up meetings helps the client to determine which form of support is better. This honor project is focusing on analyzing the data from the Pennsylvania State University Statistical Consulting Center over the past few years and it gives us a detailed picture of the general situation of the university academic statistical consulting center.

1.4 Research Objective

We have two main research objectives in this project:

The first one is to present typical features of consulting meetings and basic characteristics of clients.

The second one is to improve the quality and efficiency of the consultant's work. By combining the client's background information and the consultant's consultation summary, we are focusing on finding the relationship between two forms (consulting request form and consultation summary form) and use one side information to predict the other. In this situation, even though there is missing information in one form, we can utilize the relationship to fill in the blank. This can provide more specific guidance and instructions to help or get help efficiently for both the consultant and the client.

To be specific, the following research questions are addresses in this thesis:

Can we use the information that the client provided to predict the specific statistical concepts that the consulting meeting will focus on? Its corresponding statistical question is: is there any single English word or any combination of 2 words in consulting form text blocks has a significant relationship with consultation keywords variable in the consultant summary form?

Although there might be some differences between the industrial consulting center and the academic consulting center, their aim is the same: to provide quality statistical support as efficiently as possible. So the results shown in this project are also worthy of reference for industrial statistical consulting center managers.

Chapter 2

Exploratory Data Analysis

2.1 Data Description

The data is provided by the Statistical Consulting Center at the Pennsylvania State University. It contains consultation records of 7 semesters, from spring semester of 2015 to spring semester of 2017. The raw data set is a combined form, merged from the consulting request form and the consultant summary form. The consulting request form data is collected before the consulting meeting and it contains manually-entered text blocks and several categorical variables. As the text blocks are typed in by the client, there exist some typos and different abbreviations. The consultant summary is prepared by the consultant after the consulting meetings and it is longer and contains more information than the request form.

In the raw data set (combined form), there are twenty columns in total (shown in table 2.1). The columns above the separation line are contained in the request form and those below the separation line are contained in the summary form. Basically, there are 3 parts that the variables belong to: client personal information (name, email, department, status), project information (title, goal, data, some features) and consultant summary information (summary content, keywords, follow-up).

Column Attributes

Variable	Type	Description
Timestamp1	timestamp	time when the client submit the application form
Client Full Name	text	the full name of the client
Email	text	the email of the client
Department	text	the department the client is in
Status	text	the status the client is
Project Title	text	the project the client has question about
Research Goal	text	the research goal the client would like to achieve
Data Collected	binary	if the data for client's project is collected
Data Analyzed	binary	if the data for client's project is analyzed
Co-authorship	binary	if the data for client's project is co-authorship
Grant Potential	binary	if the data for client's project has grant potential
Remote Meeting	binary	if the data for client's project requires remote meeting
Timestamp2	timestamp	starting time of consultation meeting
Consultant Full Name	text	the full name of the consultant
follow up	binary	if this meeting is a follow up
Consultation Summary	text	the summary for this meeting written by consultant
Consultation Key Words	text	consultation summary key words
Fullow-up action	binary	if this meeting needs a follow up
Type of follow up	text	what kind of follow up this meeting needs
580	binary	if this meeting is required by course STAT580

Table 2.1: Variable name, type and description

2.2 Visualizations

There are a couple of descriptive graphs of some variables we are interested in. These graphs provide us with basic information about the characteristics of University SCC clients and the features of consulting meetings.

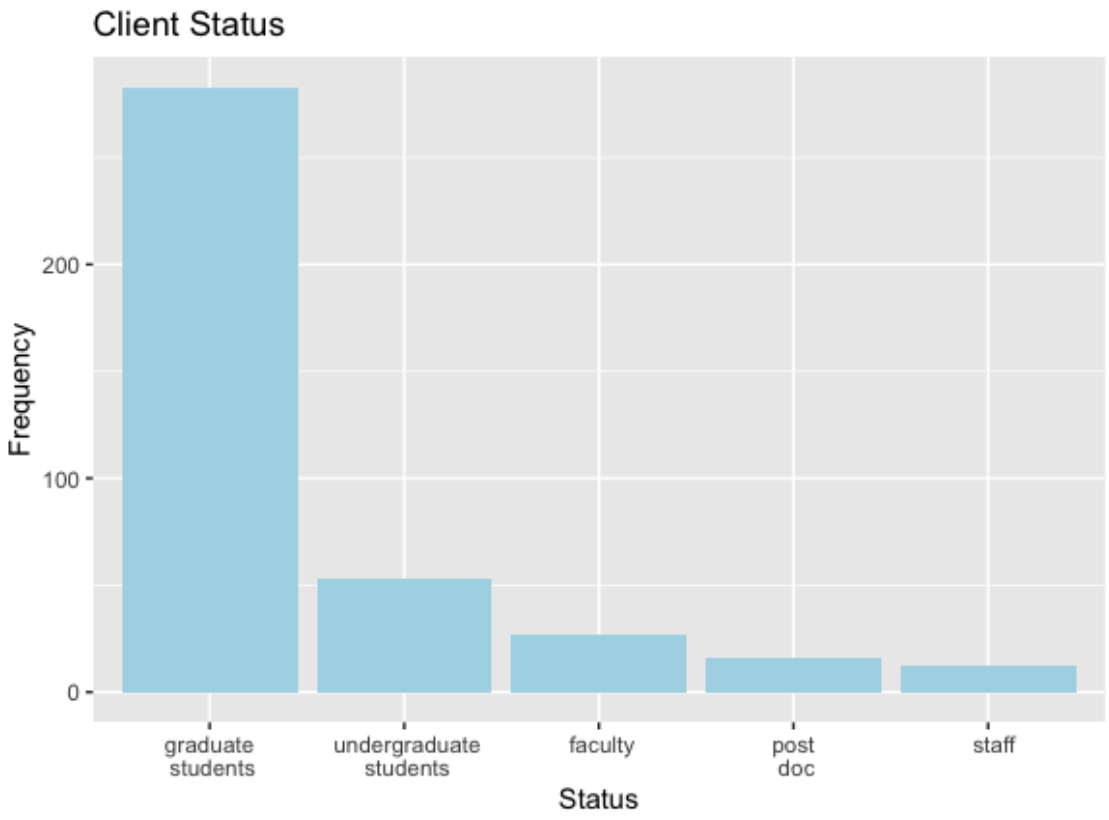

Figure 2.1: Histogram of client status

This is a histogram of counts of the client's status. In this graph, all the status listed in the graph belong to the Penn State University. As we can see, "graduate students" is the biggest group, and second is "undergraduate students", followed by "faculty", "post doc" and "staff". Among total 451 clients, 283 clients are graduate students and this amount is even bigger than the sum of the other four groups. The question that asks about the client's status on the website is a multiple choice question and it's required, the data is credible and representative.

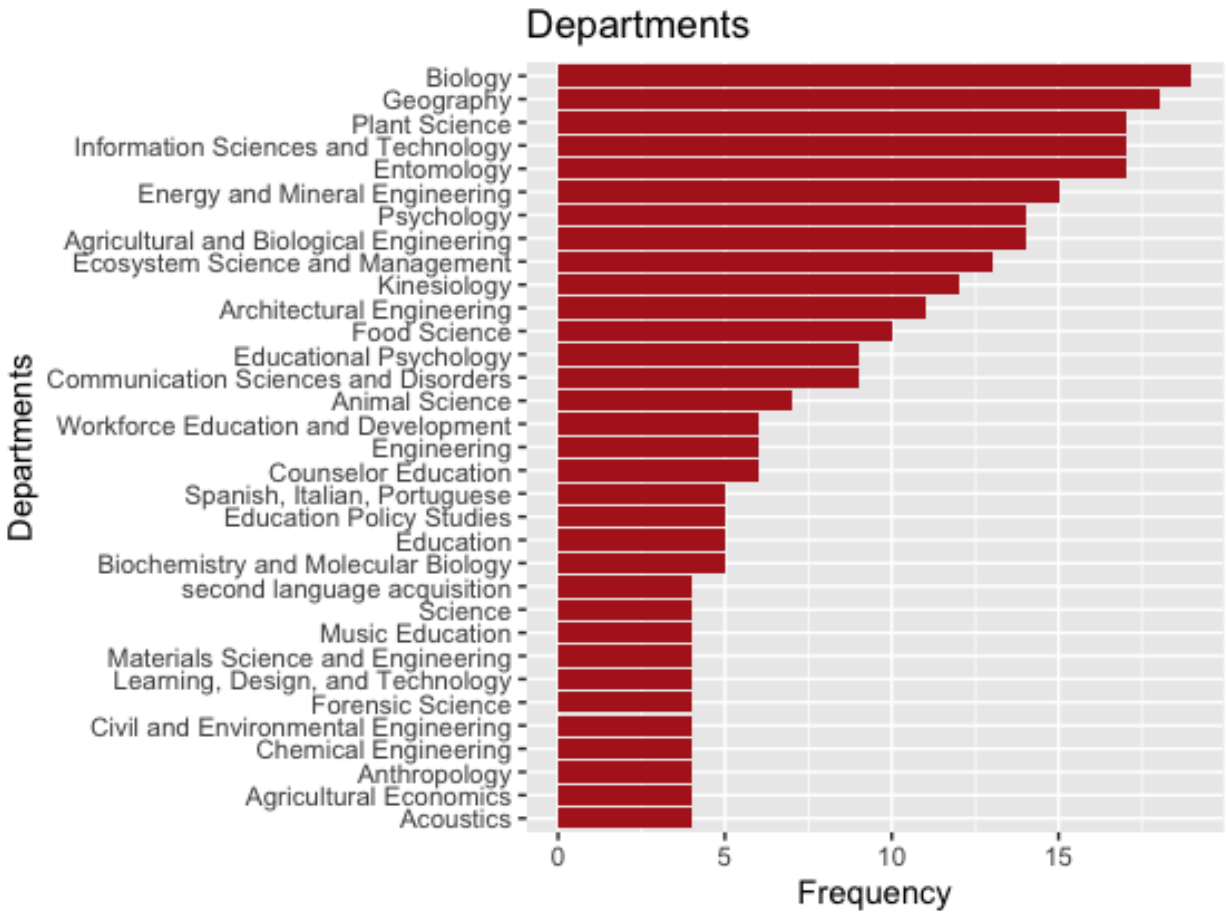

Figure 2.2: Bar chart for departments

Figure 2.2 presents a bar chart of counts of departments the client belongs to. The department data is complicated as it's typed manually. So we create another column to represent the department data after expanding the abbreviates and correcting the typos. We sort the counts in a decreasing order and we can observe that the 5 most frequent departments are "biology", "geography", "plant science", "information sciences and technology (IST)" and "entomology". We find most of the clients are in science-related departments, followed by the engineering-related departments and lastly art-related and business-related departments. One possible reason of this distribution is that there is a big difference between the sizes of researchers in different departments.

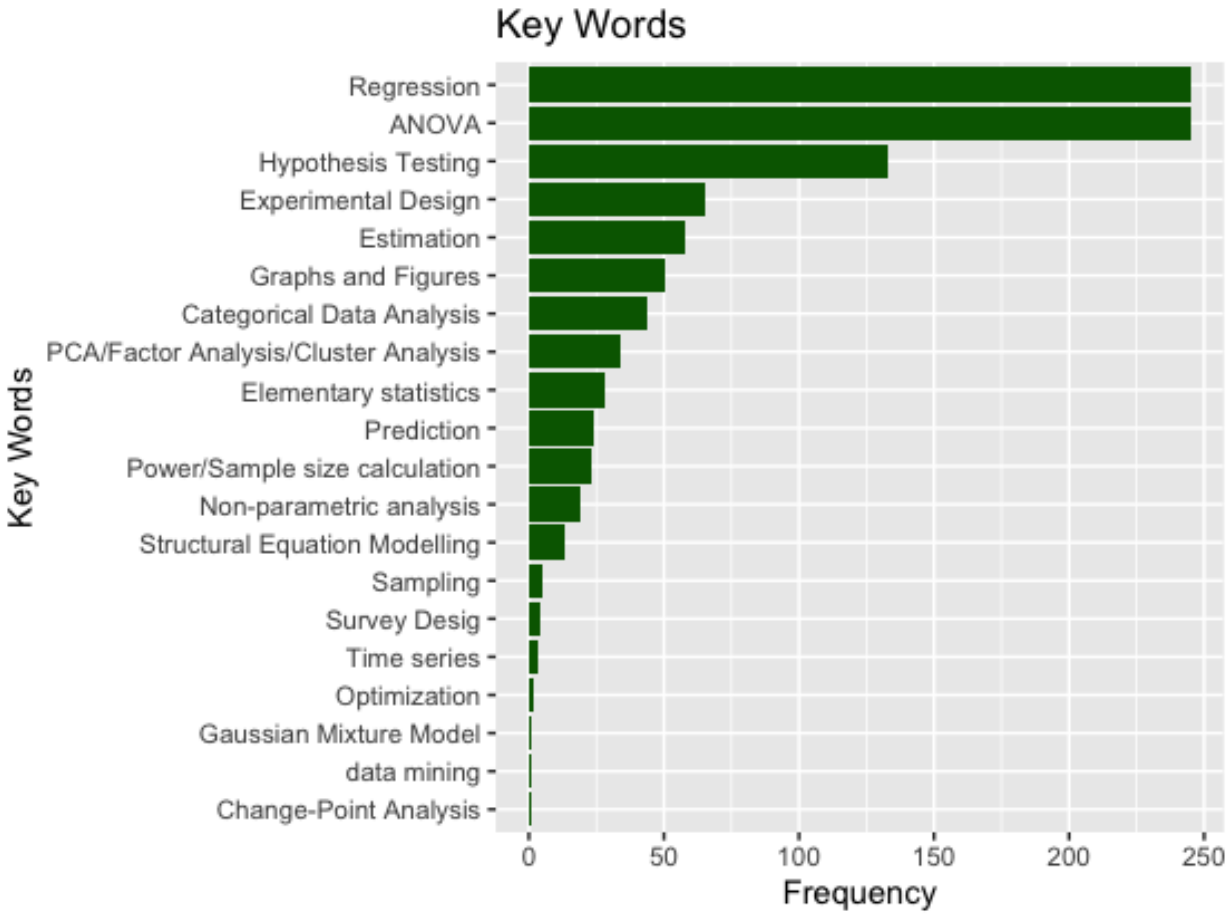

Figure 2.3: Bar chart for statistical keywords

Figure 2.3 presents a bar chart for the counts of statistics keywords that appeared in the "Consultation Key Words" column. The keyword column provides multiple choices that are selected by the consultant after the consulting meeting. We treat each choice as a binary response variable in the analysis. We count the number of presence for each keyword. Two of these keywords, "Regression" and "ANOVA", have the exact same frequency because they are tied and they always appear together.

The most frequent statistical keywords are "Regression", "ANOVA", "Hypothesis Testing", "Experimental Design" and "Estimation". The infrequent statistical concepts are "Gaussian Mixture Model", "Data mining", "Change-point analysis" and "Optimization" as they are either rarely used in research or too specific to be used on general models.

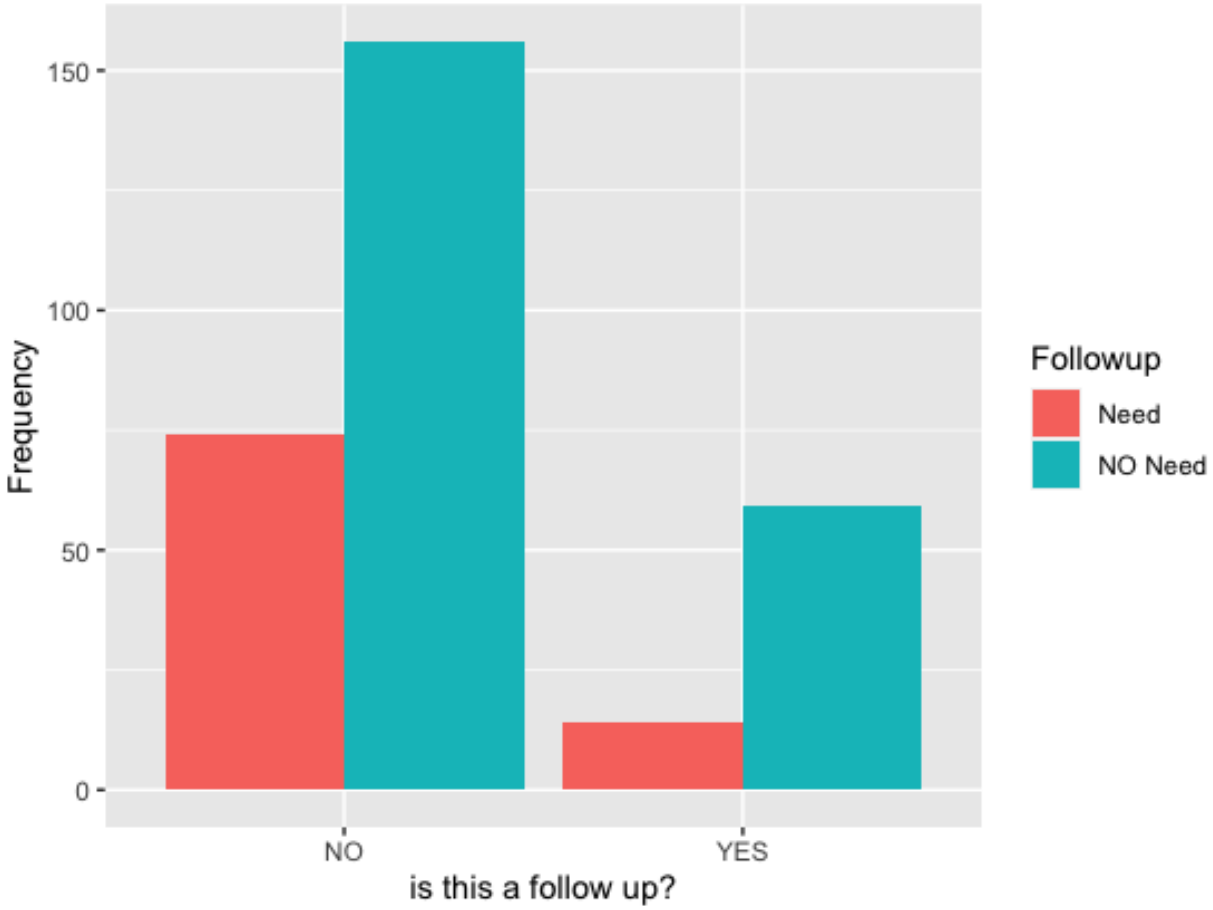

Figure 2.4: Histogram for followup

Figure 2.4 provides a histogram of the follow-up status. It is constructed by 2 variables: if this meeting is already a follow-up and after this meeting, if the client needs a follow-up. The x-axis represents if this meeting is a follow-up and the y-axis represents the frequencies of meetings. The red column represents the number of meetings that the client does need a follow-up later and the cyan column represents the number of meetings that the client does not need follow-up. We observe that when this meeting is not a follow up meeting, there are 79 clients, out of 235, need a follow-up and the follow-up rate is around 33.62%. This means that with the help of consultants, 66.38% of the clients can solve their questions completely within one consulting meeting. But if this meeting is already a follow-up meeting, 14 clients, out of 73, still need a follow-up and the follow-up rate decreases to 19.19%. So 80.82% of the clients can solve their questions in follow-up meeting(s).

Chapter 3

Token Screening

We first needed to process the text data by vectoring it. This chapter described the screening process and introduced important statistical terms that were utilized throughout this process.

3.1 Text Corpus

Text corpus is a massive collection of text and it's widely used in the data mining process. It's basically the same with "data set" in machine learning and it can be dealt as a vector.

The first step was to convert our text data to a text corpus. As we had 6 predictors: "research goal", "data collected", "data analyzed", "co-authorship", "grant potential" and "remote meeting", we put them into one text corpus so that we could did further analysis. Only "research goal" column contained the text block, and other columns were binary choices of "Yes/No".

In order to combine them with"research goal" variable, we appended some specific English words to "research goal" data based on the levels of each binary variables. Specifically, for "data collected" predictor, if a single entry was "yes" in "data collected" column, we updated the corresponding entry in "research goal" column by appending an English word "datacollected", and if the entry was "no", the corresponding entry in "research goal" stayed the same. We did the same thing for the additional 4 predictors. After we got all the entries in "research goal" updated, our text corpus was ready and we moved to the next step: tokenization.

3.2 Tokenization

Tokenization is the process of breaking down text corpus into units - tokens. The token can be in multiple types and can contain different amount of words, which is explained further in subsection 3.2.1 N-Gram. Before tokenizing our text corpus, we simply modified the tokens by removing punctuation, getting rid of stop words and numbers and converting all letters to the lower case. In this project, we included 1-gram and 2-gram tokens.

3.2.1 N-Gram

The definition of n-gram is defined as a contiguous sequence of n items from a given sample of text[3]. Basically, N-Gram is the combination of n adjacent words. For 1-gram token, it means every single word counts and for 2-gram, it means we consider the combination of 2 words such as "sample size". An example for 1-gram, 2-gram and 3-gram tokens is shown in Table 3.1.

Original Text: "Tokenization is a process of breaking down text corpus into units."

N-gram	Tokens
1-gram (unigram)	"Tokenization", "is", "a", "process", "of", "breaking", "down", "text", "corpus", "into", "unites"
2-gram (bigram)	"Tokenization is", "is a", "a process", "process of", "of creaking", "breaking down", "down text", "text corpus", "corpus into", "into units"
3-gram (trigram)	"Tokenization is a", "is a process", "a process of", "process of breaking", "of breaking down", "breaking down text", "down text corpus", "text corpus into", "corpus into units"

Table 3.1: N-gram example

3.2.2 Tokenization results for 1-gram and 2-gram

The figure below shows the 20 most frequent tokens of 1-gram and 2-gram, separately. The data in the "count" column represents the number of occurrences of the term across all text blocks and the data in "support" column represents the number of text blocks that contain the term. The terms are sorted by their values of support. The input text is the data of "research goal" variable, our text corpus.

	term	count	support
1	datacollect	339	339
2	dataan	272	272
3	data	244	148
4	use	174	121
5	na	85	85
6	analysi	96	78
7	differ	116	74
8	like	89	67
9	need	76	64
10	statist	71	62
11	determin	64	59
12	variabl	67	48
13	measur	62	48
14	help	57	47
15	two	51	45
16	studi	58	43
17	model	74	42
18	remotemeetingrequir	42	42
19	want	49	40
20	collect	47	40
⋮	(2153 rows total)		

(a) 20 most frequent 1-gram tokens

	term	count	support
1	datacollect dataan	261	261
2	na datacollect	33	33
3	dataan remotemeetingrequir	27	27
4	need help	24	22
5	data set	25	19
6	analyz data	14	13
7	collect data	14	13
8	data collect	14	13
9	make sure	13	13
10	data datacollect	12	12
11	repeat measur	11	11
12	statist analysi	11	11
13	dataan grantpotenti	10	10
14	thank datacollect	10	10
15	analysi datacollect	9	9
16	datacollect remotemeetingrequir	9	9
17	model datacollect	9	9
18	research question	9	9
19	determin best	8	8
20	like use	8	8
⋮	(7850 rows total)		

(b) 20 most frequent 2-gram tokens

Figure 3.1: Tokenization result for 1-gram and 2-gram

From the counts of 1-gram and 2-gram tokens, we could note that the number of unique 2-gram tokens was close to 4 times more than that of unique 1-gram tokens. Theoretically, this meant that the unique 2-gram tokens didn't show as frequently as unique 1-gram tokens did. The fact that the counts of 2-grams tokens are all smaller than the counts of the equally ranked 1-gram tokens confirmed this deduction.

3.3 TF-IDF

Tf-idf is an important statistic we care about during the screening process.

3.3.1 What is tf-idf?

Tf-idf is, an abbreviate of term frequency–inverse document frequency[4], a statistic representing how related the word is to a documentation in text corpus. Its value is a production of tf value and idf value. The lower tf-idf value is, the more related the word is to the documentation.

3.3.2 TF-IDF Calculation

Tf is the term frequency[4], tf(t,d), where t is term and d is document. There are many tf weights and in this project, we consider tf simply as the proportion of the term t in documentation d. The formula for tf is

$$tf = \frac{\text{Number of times that term t occurs in documentation d}}{\text{Total number of terms in documentation d}}$$

The value of tf is between 0 to 1, and 0 means the term t does not show in the documentation d at all and 1 means the documentation d only contains the term t. For example, in a string d, "The day after tomorrow is tomorrow after tomorrow",

$$\text{tf(tomorrow, d)} = \frac{\text{Number of times that term "tomorrow" occurs in string d}}{\text{Total number of terms in string d}} = \frac{3}{8}$$

Idf is the inverse document frequency[4], idf(t,D), where t is the term and D is the text corpus. The value of idf reflects how much information the term t provides. It is a logarithmic function of the ratio of total amount of documents in the text corpus to the number of documents with the term t. The formula for idf is

$$idf = \log(\frac{\text{Total number of documents}}{\text{Number of documents with word t in D}})$$

Since the value of ratio inside log is always greater than or equal to 1 and according to logarithm function properties, the value of idf is always greater than or equal to 0. If idf(t,D) is 0, it means that all documentations in the text corpus contain the term t and as idf value increases, there are fewer documentations that contain the term t, so the term t provides less information.

The tf-idf value is calculated by multiplying tf value by idf value, and the formula is

$$\text{tf-idf(t,d)} = tf(t,d) * idf(t,D) \quad \text{where d belongs to D}$$

Because the range of tf is [0,1] and the range of idf is [0, ∞], the range of tf-idf is also [0, ∞]. When the term t appears more frequently in documentation d, tf(t,d) becomes larger but it is still less than 1. When the term t appears more frequently in the text corpus, which means that the number of documentations that contain the term t becomes bigger, idf(t,D) is closer to 1. So lower the tf-idf value is, more important the term t is to documentation d in the text corpus.

3.4 Significant Words

The final step was to screen out the significant words. In other words, we filtered out the meaningful tokens among all the tokens in the text corpus. We applied tests to 1-gram tokens and 2-gram tokens separately and used the 2-sample t test and wilcoxon test in this step, which were introduced briefly below.

What is two-sample t-test?

Two-sample t-test is a commonly used hypothesis test and it is used to check if there is a significant difference between means of two groups[5]. The null hypothesis (H_0) of two-sample t-test is there is no difference between two population means ($\mu_1 = \mu_2$) and the alternative hypothesis(H_a) is there is difference between the population means ($\mu_1 \neq \mu_2$). In order to test if the difference between the 2 populations means is big enough to reject (H_0), we need to calculate t-statistic, whose

formula is:

$$t - statistic = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{{s_p}^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

where $\bar{x}_1$ and $\bar{x}_2$ are sample means; n_1 and n_2 are sample sizes; s_p is the pooled sample variance and

$$s = \sqrt{\frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2}}$$

where s_1 and s_2 are sample standard deviations.

Then we need to use t table to find what the critical value is at significant level of 0.05. If t statistic value is bigger than critical value, we say that H_0 can be rejected and the conclusion is H_a, otherwise, we say there is no difference between the 2 population means (H_0). In practice, we do not need to calculate t by hand but use R function called "t.test(x, y,..)".

What is Wilcoxon test?

Wilcoxon test is a non-parametric hypothesis test and it is used to compare means of two dependent samples[6]. The difference between the 2-sample t-test and the wilcoxon test is that to use 2-sample t-test, the data needs to satisfy normality assumption but wilcoxon test does not care about the data distribution. Besides, wilcoxon test can be applied to 2 dependent variables related to same independent variables, which is exactly the reason why we use wilcoxon test in this project. As we do not need to run wilcoxon test by hand, we won't include the calculation steps here. Instead, we use R function: "wilcox.test(x, y,..)".

3.4.1 Procedure

For each keyword, we constructed our new data set, which contained one binary response variable indicating whether the keyword had been selected by the consultant in the meeting summary report, and many continuous predictors which were the tf-idf values of tokens. The screenshot of the new data set is shown below in Figure 3.2.

response1	acoustic	also	comparative	compare	comparing	conducted	consistency
TRUE	0.2005142	0.10002198	0.2167328	0.12267780	0.14186072	0.1523553	0.2005142
FALSE	0.0000000	0.00000000	0.0000000	0.00000000	0.00000000	0.0000000	0.0000000
TRUE	0.0000000	0.00000000	0.0000000	0.00000000	0.00000000	0.0000000	0.0000000
FALSE	0.0000000	0.00000000	0.0000000	0.06815433	0.00000000	0.0000000	0.0000000
TRUE	0.0000000	0.00000000	0.0000000	0.00000000	0.13640454	0.0000000	0.0000000
FALSE	0.0000000	0.00000000	0.0000000	0.00000000	0.00000000	0.0000000	0.0000000
TRUE	0.0000000	0.00000000	0.0000000	0.26669086	0.00000000	0.0000000	0.0000000
FALSE	0.0000000	0.13160786	0.0000000	0.00000000	0.00000000	0.0000000	0.0000000
TRUE	0.0000000	0.00000000	0.0000000	0.00000000	0.00000000	0.0000000	0.0000000
TRUE	0.0000000	0.00000000	0.0000000	0.00000000	0.00000000	0.0000000	0.0000000
TRUE	0.0000000	0.00000000	0.0000000	0.00000000	0.00000000	0.0000000	0.0000000
TRUE	0.0000000	0.00000000	0.0000000	0.00000000	0.00000000	0.0000000	0.0000000
FALSE	0.0000000	0.00000000	0.0000000	0.00000000	0.00000000	0.0000000	0.0000000

Figure 3.2: Dataset sample of the screening process

We then performed the predictor screening by 2-sample t-tests. For each token, it compared the tf-idf between response = True v.s. response = False. We kept the tokens that had p-value less than 0.05 as the candidate predictors.

3.4.2 Screening result for 1-gram tokens

For each selected keyword, the distribution of counts of 1-gram significant words is shown in Figure 3.3.

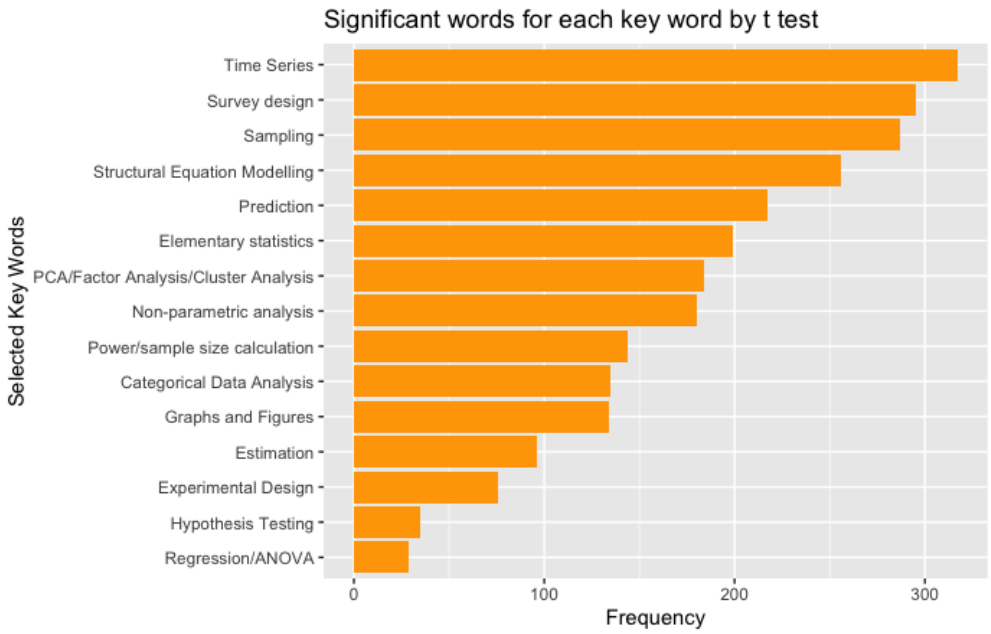

Figure 3.3: Bar chart for count of 1-gram significant words for each keyword by t test

The keyword that had most significant words was “time series”, followed by “survey design” and “sampling”. In Chapter 2.3 Visualization, we had a bar chart of counts of keywords. The most frequent keywords were “regression, “ANOVA” and “hypothesis testing”, and “time series” and “survey design” were among the 6 least frequent keywords. We suspected that more tokens have p-value < 0.05 for the low frequency response variables because the normality assumption in the t-test did not hold when the sample size was small in the rare group. So we also used the same procedure and applied the wilcoxon test for each token. We plotted the bar chart for counts of significant words and also a connected scatter plot of keywords frequencies, counts of significant words by the 2-sample t-test and counts of significant words by the wilcoxon test, shown in Figure 3.4 and Figure 3.5.

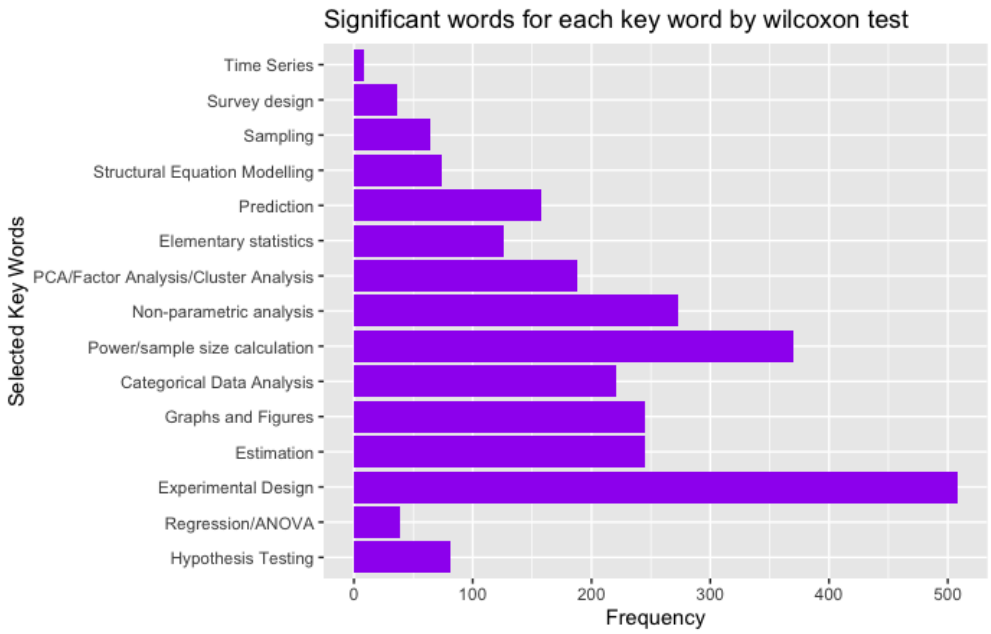

Figure 3.4: Bar chart for count of 1-gram significant words for each keyword by wilcoxon test

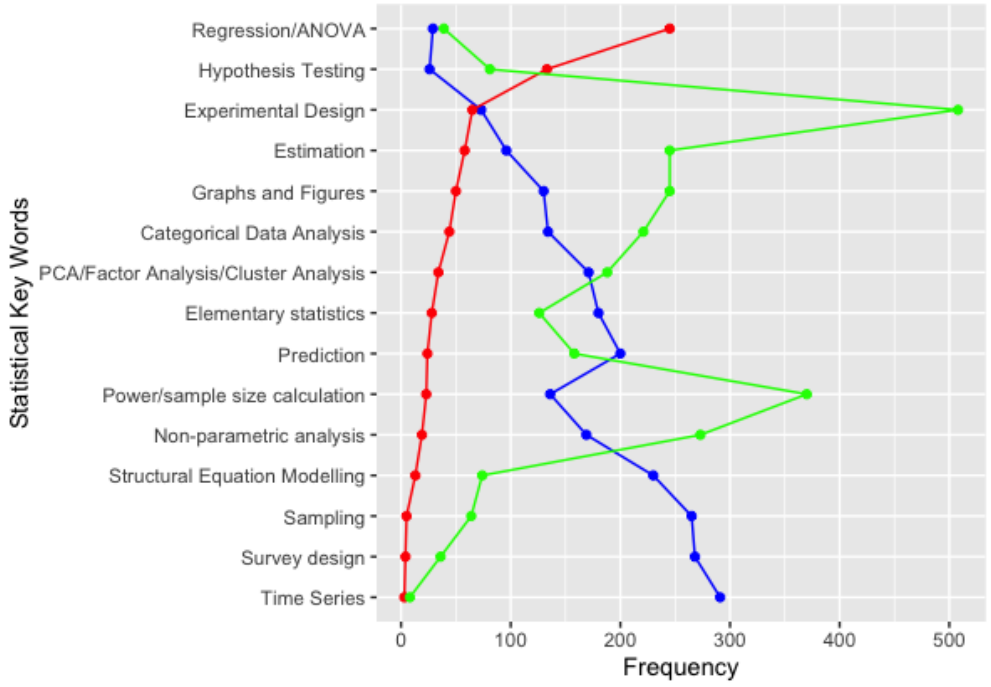

Figure 3.5: Scatterplot for counts of each keyword and counts of significant words for each keyword (1-gram)

In Figure 3.5, the red line represents the keywords frequencies, the blue line represents the counts of significant words that were screened by the 2-sample t test and the green line represents

the counts of significant words that are screened by the wilcoxon test.

From this scatter plot, we could note that for several statistical keywords, such as “experimental design”, “estimation” and “Power/sample size calculation”, there existed big difference between the result of t-test and the result of the wilcoxon test. For “Non-parametric analysis” and other less frequent statistical keywords, as the keyword count became smaller, the difference became bigger. This confirmed that for some statistical keywords, t-test result was not inaccurate, especially for these infrequent words.

3.4.3 Screening result for 2-gram tokens

For 2-gram tokens, we used the same procedure and criteria to determine significant words. As we mentioned in section 3.2.2, the number of unique 2-gram tokens was way bigger than that of unique 1-gram tokens, so theoretically, there should be more significant words among 2-gram tokens than 1-gram tokens. According to the result of 1-gram, we observed that the wilcoxon test was more appropriate than t test because for some infrequent keywords, the small data size violated normality assumption of t test. So we applied the wilcoxon test for 2-gram tokens. For each selected keyword, the distribution of counts of 2-gram significant words by wilcoxon test is shown in Figure 3.6.

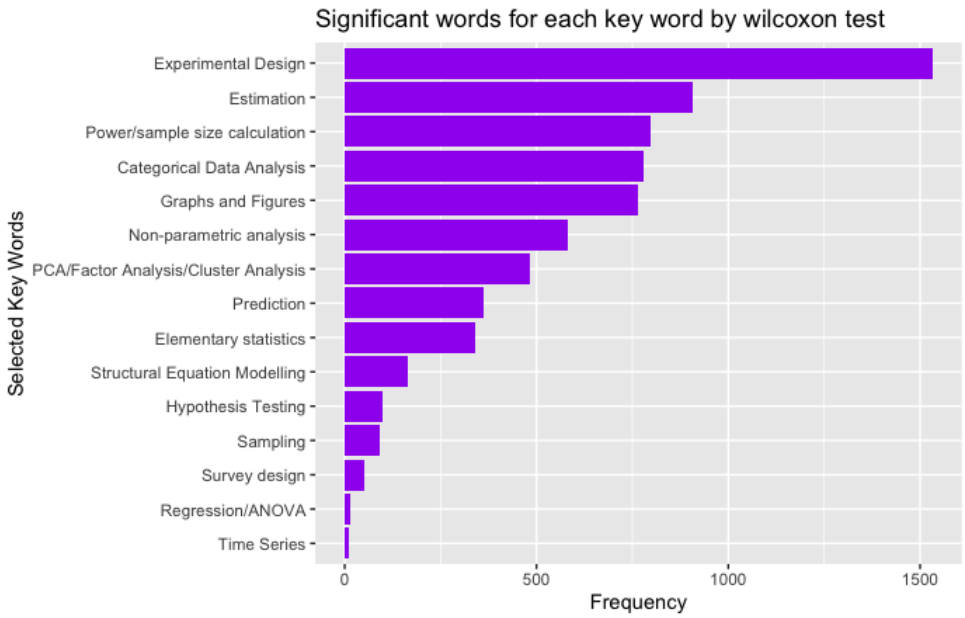

Figure 3.6: Bar chart for count of 2-gram significant words for each keyword by wilcoxon test

Chapter 4

Logistic Regression

4.1 Variable Selection

Now, we have found both 1-gram and 2-gram significant words that have a relationship with statistical keywords in the consultation summary. We noticed that for some keywords, there were too many significant words, which did not provide much information. In order to find the best subset of these significant words, we needed to apply variable selection method.

4.1.1 What is logistic regression variable selection?

The conventional linear regression, such as ANOVA, normally has a continuous response. When we have noncontinuous response, it does not fit well. In this project, our response is binary variable so the logistic regression is more appropriate.

There are several assumptions that must be satisfied before we use logistic regression. First, the response variable is dichotomous or the sum of dichotomous responses. As we have "response1" variable, which only has two levels: TRUE or FALSE, this assumption is satisfied. Second, the observations must be independent of one another. Because we assume that each word is independent of each other, this assumption is also satisfied. Last, there is no obvious outliers in our data.

4.1.2 Logistic Model

We sorted the p-values of significant words screened in Chapter 3 in an increasing order and selected the top 30 1-gram significant words and top 10 2-gram significant words. We fitted a logistic regression model with those 30 tf-idfs as candidate predictors. In the following results, "regression" was the selected keyword. The model summary of 1-gram significant words is presented in Figure 4.1.

Top 30 significant words included multiple types of words, such as nouns, adjectives, verb and even prepositions.

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.5669     0.1513  -3.746  0.00018 ***
dataanalized     3.8948     2.3165   1.681  0.09270 .
datacollected    5.8272     2.8055   2.077  0.03779 *
different        6.5052     3.3202   1.959  0.05008 .
two              7.8004     4.5642   1.709  0.08744 .
create        -155.4281 19705.9100  -0.008  0.99371
distribution   -38.5690    36.5573  -1.055  0.29141
regression       4.3827     2.7072   1.619  0.10547
used            13.4337     5.7192   2.349  0.01883 *
continuous     130.7164 24589.3473   0.005  0.99576
developing     -94.3873 13156.2714  -0.007  0.99428
measures         7.8595     4.9089   1.601  0.10936
without        309.2931 51859.3657   0.006  0.99524
effect           5.5801     3.7267   1.497  0.13431
control          6.2224     5.0284   1.237  0.21592
anova            5.3575     3.4498   1.553  0.12043
explore        154.9049 37593.1889   0.004  0.99671
minitab        100.3950 16416.9023   0.006  0.99512
exercise      -128.8998 17670.6435  -0.007  0.99418
pressure      -117.8583 20041.2156  -0.006  0.99531
sampling       -23.7901    15.0328  -1.583  0.11353
years           29.8533    19.6392   1.520  0.12849
state          254.0088 35732.5238   0.007  0.99433
differences     17.7795     9.2852   1.915  0.05551 .
rates           36.0194    21.1473   1.703  0.08852 .
survival      -105.0546 10446.3970  -0.010  0.99198
far            168.1468 30363.7802   0.006  0.99558
addition        48.7463    37.7933   1.290  0.19712
ran            381.1972 61284.4436   0.006  0.99504
behavior       101.9507 17561.9645   0.006  0.99537
```

Figure 4.1: Logistic model summary for "regression"

4.1.3 Logistic Variable Selection Process

Instead of eliminating tokens only based on their p-values, we would like to find the best combinations of those tokens. So we could not simply rely on p-value but needed other model selection statistics: Akaike Information Criterion (AIC) and Bayesian information criterion (BIC).

AIC

AIC is an abbreviate of "Akaike information criterion"[7], which is a model quality estimator for model selection. It is used to choose the best predictor subset. The mathematical equation for AIC is

$$AIC = -2logL + 2p$$

where L is likelihood function and p is number of parameters in model.

In this equation, 2p was a constant and logL was the "relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model"[8]. The lower AIC value was, the less information the model has lost. So in practice, we preferred the model with the lowest AIC.

BIC

The full name of BIC is "Bayesian information criterion"[9] and BIC is also an estimator used to evaluate the quality of a model. The mathematical equation for BIC is

$$BIC = -logL + log(n) * p$$

where L is likelihood function, p is number of parameters in model and n is data size.

We could refer to the only difference between the equations of BIC and AIC was the coefficient of p: 2 for AIC and log(n) for BIC. This meant BIC took sample size into account while AIC did not. Since BIC had a larger penalty, we chose BIC for the model selection.

BIC Sequence

When we applied the logistic regression variable selection method, the result was not a single list but a sequence of multiple lists. We chose "regression" as our selected keyword and we did this variable selection in both directions for 1-gram significant words. The starting model, step 1 model and the final model in the sequence are shown below.

```
Start:  AIC=638.12
response1 ~ dataanalized + datacollected + different + two +
    create + distribution + regression + used + continuous +
    developing + measures + without + effect + control + anova +
    explore + minitab + exercise + pressure + sampling + years +
    state + differences + rates + survival + far + addition +
    ran + behavior

               Df Deviance    AIC
- explore       1   455.79 633.02
- far           1   456.28 633.52
- pressure      1   457.39 634.63
- effect        1   457.44 634.67
- dataanalized  1   457.52 634.75
- control       1   457.84 635.08
- regression    1   457.86 635.09
- behavior      1   458.62 635.85
- exercise      1   458.86 636.09
- measures      1   459.13 636.36
- continuous    1   459.22 636.45
- two           1   459.29 636.52
- different     1   459.32 636.55
- anova         1   459.47 636.71
- minitab       1   459.61 636.84
- addition      1   459.63 636.86
- ran           1   459.70 636.93
- datacollected 1   459.79 637.02
- differences   1   460.04 637.27
- create        1   460.56 637.79
- rates         1   460.80 638.03
<none>              454.77 638.12
- developing    1   461.24 638.47
- without       1   461.46 638.69
- distribution  1   461.73 638.96
- years         1   461.78 639.01
- sampling      1   462.14 639.37
- used          1   463.01 640.25
- survival      1   464.09 641.32
- state         1   464.30 641.53
```

Figure 4.2: Starting Model in BIC sequence

```
Step:  AIC=633.02
response1 ~ dataanalized + datacollected + different + two +
    create + distribution + regression + used + continuous +
    developing + measures + without + effect + control + anova +
    minitab + exercise + pressure + sampling + years + state +
    differences + rates + survival + far + addition + ran + behavior

               Df Deviance    AIC
- far           1   457.26 628.38
- effect        1   458.44 629.56
- pressure      1   458.45 629.57
- dataanalized  1   458.53 629.65
- control       1   458.82 629.94
- regression    1   458.86 629.98
- behavior      1   459.62 630.74
- exercise      1   459.96 631.08
- continuous    1   460.24 631.36
- different     1   460.33 631.45
- anova         1   460.46 631.58
- two           1   460.52 631.65
- measures      1   460.58 631.71
- minitab       1   460.62 631.74
- ran           1   460.71 631.83
- datacollected 1   460.78 631.90
- differences   1   461.00 632.12
- create        1   461.61 632.74
- rates         1   461.90 633.02
<none>              455.79 633.02
- without       1   462.43 633.55
- developing    1   462.43 633.56
- years         1   462.77 633.89
- sampling      1   463.17 634.29
- distribution  1   463.22 634.34
- addition      1   464.15 635.27
- used          1   465.10 636.22
- survival      1   465.11 636.23
- state         1   465.29 636.41
+ explore       1   454.77 638.12
```

Figure 4.3: Step 1 Model in BIC sequence

```
Step:  AIC=604.86
response1 ~ datacollected + distribution + used + developing +
    measures + exercise + sampling + years + state + differences +
    rates + survival + addition + continuous

               Df Deviance    AIC
<none>              513.19 604.86
+ without       1   507.40 605.19
+ minitab       1   507.50 605.28
+ anova         1   508.05 605.83
- rates         1   520.27 605.83
+ regression    1   508.25 606.04
- exercise      1   520.66 606.22
+ behavior      1   508.45 606.24
+ create        1   508.58 606.36
+ different     1   508.66 606.45
+ control       1   508.68 606.47
+ effect        1   508.74 606.52
- survival      1   521.27 606.83
- continuous    1   521.42 606.98
+ two           1   509.22 607.01
- measures      1   521.47 607.03
- sampling      1   521.56 607.12
- years         1   521.93 607.49
- state         1   522.10 607.66
+ ran           1   510.08 607.86
- developing    1   522.40 607.96
+ far           1   510.40 608.18
+ pressure      1   510.64 608.42
- used          1   522.88 608.44
- addition      1   522.92 608.48
+ dataanalized  1   510.88 608.66
- differences   1   523.14 608.70
- distribution  1   523.57 609.13
+ explore       1   512.26 610.04
- datacollected 1   524.64 610.20
```

Figure 4.4: Final Model in BIC sequence

4.2 Numerical Results for 1-gram Tokens

Among all the reduced models in the sequence, we chose the model with the smallest BIC value as our final model and the combination of its factors was the best subset. The final model was:

$$response1 - -0.1783 + 13.8369 * used - 124.1753 * exercise + 7.7319 * datacollected$$

$$-90.7067 * developing - 27.4823 * sampling + 27.9124 * years - 100.4319 * survival$$

$$+55.0771 * addition - 50.1714 * distribution + 11.8769 * measures + 21.4700 * differences$$

$$+221.9378 * continuous + 233.6747 * state + 36.6139 * rates$$

with BIC value = 604.86

The tables below shows the logistic models of 1-gram significant words screened by t-test and wilcoxon test, seperately, and their BIC values.

Logistic Models Table for significant words screened by t test

Selected predictor	Logistic Model	BIC value
regression/ANOVA	-0.1783 + 13.8369*used - 124.1753*exercise + 7.7319*datacollected - 90.7067*developing - 27.4823*sampling + 27.9124*years - 100.4319*survival + 55.0771*addition - 50.1714*distribution + 11.8769*measures + 21.4700*differences + 221.9378*continuous + 233.6747*state + 36.6139*rates	604.86
Hypothesis Testing	-0.8571 - 11.3819*model - 271.0710* point + 124.0445 *blood + 114.7808*sense	542.85
Experimental Design	-1.7719 + 4.9233*design -12.2881* remotemeetingrequired	376.82
Estimation	-1.7551 - 13.8069*help - 932.649*can	351.13
Graphs and Figures	-1.8939 - 487.5389*project - 16.8228*need	319.44
Categorical Data Analysis	-2.1454 - 496.1926*model	293.98
PCA	-2.4243 - 492.2760*design	248
Elementary statistics	-2.6000 - 358.8775*help	215.86
Prediction	-2.7983 - 517.4358*design	195.97
Power/Sample size calculation	-2.3928 - 41.7478*dataanalized	182.16
Non-parametric analysis	-3.0217 - 543.5102*variables	165.96
Structural Equation Modelling	-3.370e+00 - 9.114e+02*using	126.28
Sampling	-3.54 - 137.896*datacollected	61.83
Survey Design	-3.296e+00 - 6.7e+03*datacollected	46.74

Table 4.1: 1-gram:Selected predictor, model and BIC by t-test

Logistic Models Table for significant words screened by wilcoxon test (1-gram only)

Selected predictor	Logistic Model	BIC value
regression/ANOVA	0.2664 -136.215*source -21.5484* create -372.3343*increase -91.5722*developing + 214.9919*behavior -80.7225*gaussian -353.3919*criteria -11.8039*grantpotential -125.825*exercise + 7.1843*regression + 17.126*measures -202.9480*cox	612.68
Hypothesis Testing	-1.042 + 28.149*accuracy+ 255.466*column + 122.855*sense + 79.733*selfefficacy + 16.637*blood + 446.62*inoculated	540.8
Experimental Design	-2.1558 + 174.4498*pilot + 16.0185*small + 142.6148*source + 149.8085*low + 90.9553*applications + 9.8412*temperature + 89.4136*optimize	345.55

Table 4.2: 1-gram: Selected predictor, model and BIC by wilcoxon test

Continued table (1-gram only)

Estimation	-2.3679 + 300.0463*unit + 21.0662*exercise + 20.5245*sense + 11.8455*temperature + 300.3619*membrane + 13.6104*application + 165.7908*therefore + 52.0035*propensity	323.55
Graphs and Figures	-2.5693+ 188.8368*regulate + 104.6207*tables + 5.58*smoothing + 7.0356*gaussian + 524.2702*suited + 235.9578*day + 30.6747*terms + 238.3133*principal+ 149.9059*occurs	290.15
Categorical Data Analysis	-2.7240 + 25.7958*weeks + 27.3856*pilot +413.3336*objective + 27.113*characteristics + 73.7341*satisfaction + 150.5406 *either + 119.8068 *demographics + 418.1481*fungus	261.57
PCA	-2.6713 + 175.4119*clusters + 24.7442*likert	235.08
Elementary statistics	-2.7515 +44.7457*degree	216.49
Prediction	-3.3078 +45.2834*admissions + 55.7334*failure + 20.3657*demographics +36.2811*required + 44.4937*density	177.58
Power/Sample size calculation	-3.972e+00 +1.789e+02*covered +1.311e+02*northern + 1.006e+03*resource + 1.969e+02*virus + 5.697e+01*power + 1.585e+01*farmers	133.62
Non-parametric analysis	-4.137e+00 + 7.123e+01*advisor +3.180e+01*nonparametric + 4.190e+01*anovas + 3.846e+02*membranes + 1.006e+02*emissions + 7.542e+01*boxcox + 4.248e+02*investigates + 5.841e+02*infection	128.67
Structural Equation Modelling	-4.29 +268.579*business + 155.186*sem + 44.029*drinking + 61.641*aqueous	94.12
Sampling	-5.4072 + 80.5585*seemenligy + 271.3548*attitudes + 59.3589*listservs	50.08
Survey Design	-38.17 + 127.74*talk + 78.13*tool +225.81*suitable+ 274.81*buildings	30.56

Table 4.3: 1-gram:Selected predictor, model and BIC by wilcoxon test (continued)

We could see that BIC values calculated by different tests were different. After comparing the BIC values in these two tables, we found that among all the keywords, only “regression/ANOVA” and “Elementary statistics” had smaller BIC values in the t-test table. All the other keywords had smaller BIC value in the wilcoxon test table. We wanted the model with smaller BIC values. So obviously, logistic models of 1-gram significant words screened by wilcoxon test were more

preferable. This result also confirmed our conclusion in Chapter3: “for some statistical keywords, t-test result is not inaccurate, especially for these infrequent words”.

4.3 Numerical Results for 1-gram and 2-gram Tokens

So far, the models were constructed by 1-gram tokens only. So we applied logistic variable selection on the combination of 1-gram and 2-gram significant words to see if it gave us better models. Here, the 2-gram significant words were screened by wilcoxon test.

Logistic Models Table for significant words screened by wilcoxon test (1-gram 2-gram)

Selected predictor	Logistic Model	BIC value
regression/ANOVA	-0.1366 +5.3581*dataanalized + 17.5411*measure + 114.2637*behavior + 765.6027*two differ -181.64*create + 7.4506*regression+ 402.3712*want make -10.7261*grantpotential -90.7578*developing -198.6221*cox + 187.8073*determine best + 227.9899*variable datacollect	604.52
Hypothesis Testing	-1.1417 + 37.6024*accuracy 6.4222*data datacollect + 135.8831*need determine + 80.1504*selfefficacy + 256.887*column + 105.5234*honey + 171.9195*blood pressure + 492.881*rate scale + 115.3293*survey student	533.65
Experimental Design	-2.2152 +10.1007*temperature + 157.9903*low + 333.1637*will evaluate + 151.3124*source + 94.319*optimize + 18.9668*small + 341.5419*experiment design	334.31
Estimation	-2.3679 + 300.0463*unit + 21.0662*exercise + 20.5245*sense + 11.8455*temperature + 300.3619*membrane + 13.6104*application + 165.7908*therefore + 52.0035*propensity	323.55
Graphs and Figures	-2.5693+ 188.8368*regulate + 104.6207*tables + 5.58*smoothing + 7.0356*gaussian + 524.2702*suited + 235.9578*day + 30.6747*terms + 238.3133*principal+ 149.9059*occurs	290.15

Table 4.4: 1-gram and 2-gram: Selected predictor, model and BIC by wilcoxon test

Continued table (1-gram and 2-gram)

Categorical Data Analysis	-3.0348 +43.8529*alcohol + 29.7858*pilot + 78.6635*satisfaction + 19.9331*ztest + 28.1363*weeks + 111.6256*damage + 26.7196*fungus + 160.3341*either + 128.2754*demographics + 29.4234*characteristics + 440.7519*objective + 55.8128*abuse child + 28.9245*also want + 406.4592*applied data	258.18
PCA	-2.6713 + 175.4119*clusters + 24.7442*likert	235.08
Elementary statistics	-3.157 +99.4404*fuzzy + 15.2985*analysis appropriate + 137.6867*cohort + 137.6867*clone + 53.5448*degree + 267.7241*actually + 114.7389*abroad specif + 84.1419*amount pcr + 30.597*among depend + 84.1419*agent subsample	217.63
Prediction	-3.3487 +50.1802*admissions + 61.7603*failure + 57.9002*emotional + 285.041*written + 81.0603*addit collect + 100.3604*adhere itar + 92.6404*among specific + 150.5406*addit inform	185.99
Power/Sample size calculation	-4.218e+00+ 2.126e+02*covered -1.660e+03*sample + 2.086e+03*assist determine +6.557e+01*power +4.599e+02 *consider + 1.446e+01*farmers + 1.085e+03*resource +3.571e+03*virus+ 2.182e+03*northern tanzania	133.14
Non-parametric analysis	-4.478e+00 + 4.563e+02*investigates + 4.186e+02*membranes + 7.988e+01*boxcox + 3.411e+01*nonparametric + 1.065e+02*emissions + 1.287e+02*acoustic method +7.544e+01*advisor + 6.201e+02*infection + 1.109e+02*across cell + 4.438e+01*anovas	161.64
Structural Equation Modelling	-4.29 +268.579*business + 155.186*sem + 44.029*drinking + 61.641*aqueous	94.12
Sampling	-5.4072 +59.3589*listservs + 80.5585*seemenligy + 271.3548*attitudes	50.08
Survey Design	-26.57 +225.54*buildings + 49.03*analysis tool + 176.51*appropriate research + 88.25*collect etc	30.56

Table 4.5: 1-gram and 2-gram: Selected predictor, model and BIC by wilcoxon test (continued)

By comparing the models of 1-gram significant words only (mentioned as“1-gram model” in the following) and the models of the combination of 1-gram and 2-gram significant words (men-

tioned as “combination model” in the following), we could see that most combination models were better than 1-gram models, except the models of keyword “Elementary statistics”. Among all the keywords, “Estimation”, “Graphs and Figures”, “PCA”, “Structural Equation Modelling” and “Sampling” had exactly same model no matter we used 1-gram tokens or the combination of 1-gram and 2-gram tokens. This meant that for these keywords, none of 2-gram tokens was included in their final models and their best predictor subset was constructed only by 1-gram tokens.

Chapter 5

Conclusion

This project aimed to find the relationship between the client information and consultant summary so that we could use the client information to predict which statistical topic the consulting meeting would be focusing on in the future.

This thesis provided multiple graphs that described the typical features of the clients, which gave us more detailed information about the University Statistical Consulting Center. During the process of analyzing the data, we vectored the text corpus into 1-gram and 2-gram tokens, screened them to significant words by applying the 2-sample t-test and the wilcoxon test. Then in order to find the best subset of these significant words, we applied logistic variable selection method. We fitted logistic variable selection method on 1-gram tokens and the combination of 1-gram and 2-gram tokens separately to find the best models. The outline of the entire research process is presented in Figure 5.1.

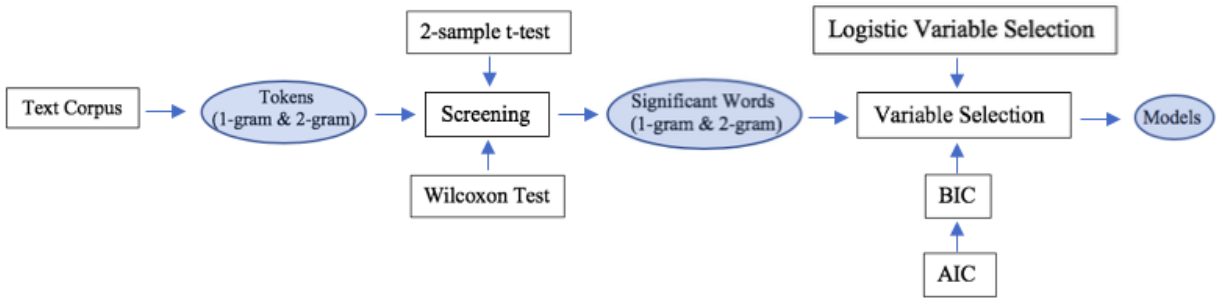

Figure 5.1: Outline

We did two comparisons in this research: 2-sample t-test v.s. wilcoxon test in the screening process and logistic model for 1-gram tokens only v.s. logistics model for a combination of 1-gram and 2-gram tokens in logistic regression.

2-sample t-test v.s. wilcoxon test

By comparing the screening results, we found that the wilcoxon test was more appropriate than the 2-sample t-test. Because in our data set, there were some infrequent keywords that had small sample size, the 2-sample t-test was inaccurate in this situation but the wilcoxon test was not.

Besides, we found an interesting phenomenon about the relationship between the frequencies of keywords and counts of their significant words. As Figure 3.5 showed, for the significant words screened by t-test, there was a negative relationship between keywords frequencies and their significant words counts. The more frequent keyword was, the fewer significant words the keyword had. However, for significant words screened by wilcoxon test, there was no clear relationship between them. But there was a trend that when the frequencies of keywords were smaller than a certain value, as the keyword became more infrequent, its significant words became fewer. More interesting was that for the most frequent keyword and the least frequent keyword listed in Figure 3.5, they had the same counts of their significant words.

1-gram tokens v.s. the combination of 1-gram and 2-gram tokens

By comparing the logistics model for 1-gram tokens only and the logistics models with the combination of 1-gram and 2-gram tokens, we discovered that adding 2-gram tokens was helpful. For most of the statistical key words, their logistics models for both 1-gram and 2-gram tokens had smaller BIC values than logistics models for 1-gram tokens only, we could say that logistic models for both 1-gram and 2-gram tokens had high qualities. So it did improve our logistic models but how much it could improve was not predicable.

Bibliography

[1]List of university statistical consulting centers. (2020, March 15). Retrieved April 26, 2020, from https://en.wikipedia.org/wiki/List of university statistical consulting centers

[2]Statistics Consulting Center. (n.d.). Retrieved March 10, 2020, from https://scc.stat.psu.edu/

[3] "N-Gram." Wikipedia, Wikimedia Foundation, 7 Apr. 2020, en.wikipedia.org/wiki/N-gram.

[4] Tf–idf. (2020, April 15). Retrieved April 20, 2020, from https://en.wikipedia.org/wiki/Tf–idfTerm frequency

[5] Unpaired Two-Samples T-test in R. (n.d.). Retrieved April 10, 2020, from http://www.sthda.com/english/wiki/unpaired-two-samples-t-test-in-r

[6] Stephanie. (2019, August 26). Wilcoxon Signed Rank Test: Definition, How to Run. Retrieved April 10, 2020, from https://www.statisticshowto.com/wilcoxon-signed-rank-test/

[7] Datalab, A. (2019, January 7). Akaike Information Criterion(AIC). Retrieved April 3, 2020, from https://medium.com/@analyttica/akaike-information-criterion-aic-7a4b58bce206

[8] AIC VS. BIC. (n.d.). Retrieved April 3, 2020, from https://www.methodology.psu.edu/resources/AIC-vs-BIC/

[9] Bayesian information criterion. (2020, April 25). Retrieved April 3, 2020, from https://en.wikipedia.org/wiki/Bayesian information criterion

Yiyang Wang

(267)584-5805 • yxw280@psu.edu

EDUCATION

Schreyer Honors College at Pennsylvania State University Graduation Date: May 2020

Computing Statistics (B.S.) and System Analysis (B.S.)

Honor: Dean's List from freshman to present

Relevant coursework: Hypothesis Testing, Regression, Probability Theory, Analysis of Variance, Survey Sampling, Calculus(I,II,III), Matrix, Linear Algebra, SAS, Combinatorics, Real Analysis, Numerical computation, Linear programming, C++ Programming, Python Programming(I,II), Java programming.

AWARD

Women in Math Research Scholarship April 2019

DataFest 2019 Finalist and Best Visualization Award April 2019

WORK EXPERIENCE

Penn State Learning University Park

Peer math tutor June 2018 – December 2019

- Helped tutees with their coursework, as well as with establishing short- and long-term study goals
- Helped tutees review material for quizzes and tests; provided useful test taking tips

Math 034 course: Math of Money University Park

Teaching assistant January 2019 – Present

- Answered students' questions about the course setting and course materials; provided additional advice and guidelines
- Held office hours for students; provided homework help and aided with exam preparation, either one-on-one or in small groups
- Collected student feedback throughout the course and communicated it to the professor

Shiny App Program University Park

Full-time employee May 2019 – December 2019

- Developed R Shiny apps individually using abundant MathJax and then tested them in upper division statistics course
- Collaborated with partner to modify previous apps by adding important features and styling them using CSS/HTML
- Presented in Shiny app showcase and exhibition for Statistics Department faculty

ACADEMIC RESEARCH

Research on Hypergeometric Functions in Geometric Scattering Theory University Park

Research Assistant, Supervised by Dr. Jeffrey Case February 2019 – January 2020

- Individual project on hypergeometric functions and their role in geometric scattering theory
- Did calculations on Heun functions and fractional Q-curvature

Research on analysis of text mining in academic statistical consulting data University Park

Research Assistant, Supervised by Dr. Le Bao November 2019 –Present

- Honor thesis project focused on analyzing multiple-typed dataset
- Screened words from inputs in text box form to 1-gram and 2-gram tokens
- Generated the best subset of tokens to predict the presence of the specific word by logistic variable selection method

SKILLS

- Computer programing skills: Python, C++, Java
- Statistics and mathematical software: Minitab, R, SAS, R shiny, Mathematica, MATLAB.
- Basic computer skills: Word, Excel, PowerPoint