

Statistics in the Wild: How Practitioners Choose Statistical Procedures

Master's Thesis
submitted to the
Media Computing Group
Prof. Dr. Jan Borchers
Computer Science Department
RWTH Aachen University

by
Yue Hu

Thesis advisor:
Prof. Dr. Jan Borchers

Second examiner:
Prof. Dr. Ulrik Schroeder

Registration date: 09.07.2019
Submission date: 21.08.2019

Eidesstattliche Versicherung

Name, Vorname

Matrikelnummer

Ich versichere hiermit an Eides Statt, dass ich die vorliegende Arbeit/Bachelorarbeit/
Masterarbeit* mit dem Titel

selbständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt. Für den Fall, dass die Arbeit zusätzlich auf einem Datenträger eingereicht wird, erkläre ich, dass die schriftliche und die elektronische Form vollständig übereinstimmen. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Ort, Datum

Unterschrift

*Nichtzutreffendes bitte streichen

Belehrung:

§ 156 StGB: Falsche Versicherung an Eides Statt

Wer vor einer zur Abnahme einer Versicherung an Eides Statt zuständigen Behörde eine solche Versicherung falsch abgibt oder unter Berufung auf eine solche Versicherung falsch aussagt, wird mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft.

§ 161 StGB: Fahrlässiger Falscheid; fahrlässige falsche Versicherung an Eides Statt

(1) Wenn eine der in den §§ 154 bis 156 bezeichneten Handlungen aus Fahrlässigkeit begangen worden ist, so tritt Freiheitsstrafe bis zu einem Jahr oder Geldstrafe ein.

(2) Straflosigkeit tritt ein, wenn der Täter die falsche Angabe rechtzeitig berichtigt. Die Vorschriften des § 158 Abs. 2 und 3 gelten entsprechend.

Die vorstehende Belehrung habe ich zur Kenntnis genommen:

Ort, Datum

Unterschrift

I hereby declare that I have created this work completely on my own and used no other sources or tools than the ones listed, and that I have marked any citations accordingly.

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie Zitate kenntlich gemacht habe.

Aachen, August 2019

Yue Hu

Contents

Abstract	xi
Acknowledgements	xiii
Conventions	xv
1 Introduction	1
2 Related work	5
2.1 Selection of Statistical Tests	5
2.2 Q&A Websites	6
3 Methodology	9
3.1 Motivations	9
3.2 Interviews	11
3.2.1 Participants	12
3.2.2 Data Collection	14
3.2.3 Data Analysis	15
3.3 Content Analysis of Q&A Sites	18
3.3.1 Websites Analyzed	18
3.3.2 Data Collection	19
3.3.3 Data Analysis	20
4 Findings	23
4.1 Interview Findings	23

4.1.1	Theme 1: Problems in Methods Selection	23
4.1.2	Theme 2: Coping Strategies	25
4.1.3	Theme 3: Recourse Utilization	30
4.2	Findings: Content Analysis of Q&A Websites	37
4.2.1	Type of Questions	37
4.2.2	Information in Questions	41
4.2.3	Problems in Questions Formulation	51
4.2.4	When Do Data Scientists Choose Statistical Tests?	64
5	Discussion	69
5.1	Possible Causes of Uncertainty	69
5.2	Reasons for the Subjective Attitude and Existing Problems	71
5.3	Types of Questions About Statistical Test Selection	74
5.4	Reasons for Vague and Missing Information	75
5.5	Proper Question Timing	78
5.6	Potential Solutions	79
6	Conclusion	81
A	Informed consent form	85
B	Interview Protocol	87
B.0.1	Interviewee Background	87
B.0.2	Walkthrough a prior task	88
B.0.3	Summary question	89
	Index	99

List of Figures

3.1	Coding process with MAXQDA 2018	17
3.2	Main code matrix	17
4.1	Software tag on ResearchGate (R37)	49
4.2	Software tag on CrossValidated (C14)	49
4.3	Answer with software code scripts on Cross-Validated (C12)	50
4.4	The process of respondents handling missing information	53
4.5	Respondent assume the missing information(C6)	55
4.6	Problem with table structure in the forum (C7)	58
4.7	The wrongly used chart on the websites (C30)	62
4.8	The process of respondents handling unclear information	63
4.9	The process of a research study	64

List of Tables

3.1	Interview participant profile	14
4.1	Information of Q&A Websites summarized from interviews	36
4.2	Types of information that askers provided in the questions	41
4.3	Types of summary statistics that askers pro- vided in the questions	43
4.4	Types of assumptions that askers provided in the questions	44
4.5	Types of assumptions that askers provided in the questions	45
4.6	Question without dataset	46
4.7	Summary of missing information	52
4.8	Summary of unclear information	56

Abstract

Statistical tests are mathematical tools for analyzing quantitative data generated in a research study. Although there are many available resources to help, several studies have indicated that selecting the appropriate statistical method is still a recognized problem among practitioners. The goal of this study were to better understand this problem and identify how do practitioners search for statistical analysis procedure. To achieve these goal, interviews were conducted to explore data scientists' concern and experiences of statistical tests selection processes. Interview results showed that practitioners are uncertain and made mistakes in selecting the appropriate statistical method, they prefer using most commonly used methods in their research domain and use available resources to help them make decisions.

Since all interview participants mentioned they have used the information on Q&A sites, we want to analyze more of how data scientists provide and ask for information on the Q&A community, what potential problems they have and how experts answer their questions. We achieved these goals through content analysis with 101 questions asking for selecting statistical methods posted in StackOverflow, CrossValidated, and ResearchGate. The results centers around four main question themes: (1) What type of questions do data scientists ask? (2) What information do data scientists provide in the questions? (3) What problems do data scientists have when providing information? (4) When data scientists choose statistical tests?

The findings got from interviews and Q&A sites analysis enrich our understanding of problems in statistical test selection and practitioners statistical information behavior. Based on the findings, design implications were proposed for the continued evolution of statistical Q&A platforms for supporting data scientists' statistical tests selection.

Acknowledgements

First of all, I would like to extend my sincere gratitude to my supervisor, Krishna Prasath Subramanian, for his instructive advice and useful suggestions on my thesis. I am deeply grateful of his help in the completion of this thesis.

Secondly, I would like to thank all people from Media Computing Group, who helped and gave me suggestions during my thesis to finish my study.

Finally, I am indebted to my parents and friends for their continuous support and encouragement.

Conventions

Throughout this thesis we use the following conventions.

Text conventions

Definitions of technical terms or short excursus are set off in coloured boxes.

EXCURSUS:

Excursus are detailed discussions of a particular point in a book, usually in an appendix, or digressions in a written text.

Definition:

Excursus

The whole thesis is written in American English.

Chapter 1

Introduction

Selection of the appropriate statistical method is a very important step in analysis of both research and industry data. A statistical test provides a mechanism for making quantitative decisions about a process or processes. The intent is to determine whether there is enough evidence to "reject" a conjecture or hypothesis about the process. For example, in the field of psychology, statistical tests of significances like t-test, z test, f test, chi square test, etc., are carried out to test the significance between the observed samples and the hypothetical or expected samples. For example, if a researcher wants to conduct a statistical test upon the significant difference between the IQ levels of two college students, then the researcher can perform the t statistical test for the difference of the two samples. A wrong selection of the statistical method not only creates some serious problems during the interpretation of the findings but also affects the conclusion of the study. In statistics, for each specific situation, statistical methods are available to analysis and interpretation of the data.

Statistical tests in psychology

There are many previous research show issues in statistical test selection and confirm this is a challenging task for practitioners in different areas [12, 51]. A variety of research also has focused on understanding uncertainty in statistical analysis and how analysts work with uncertain data, including defining typologies, task taxonomies, coping strategies, and creating analysis and sense-making models. Many scientists indicate that although a variety of studies and materials [50, 45] try to give suggestions on

Studies have indicated issues in statistical test selection

	<p>choosing the proper statistical tests, there is still no one recognized standard that everyone can use in practice. There are so many uncertainties and complexity exist in the procedure of statistical tests selection, how do practitioners deal with these difficulties and search useful information helping them make a proper decision?</p>
<p>Inadequate research in data scientists information-seeking behavior</p>	<p>This important research direction is still blank today, a better understanding of data practitioners' problems, information and resources seeking behavior can help them optimize the process of selecting statistical tests and propose a helpful solution.</p>
<p>Interview findings centers around three main themes</p>	<p>This study identifies how data scientists search statistical analysis procedure. Our starting point in this study is talking with data practitioners and understanding their concerns in a general perspective. The analysis of the interview transcripts reveals findings structured around three main themes: (a) data scientists face problems in selecting the appropriate statistical analysis method, (b) selecting statistical method is often subjective, (c) data scientists use available resources to help them make decisions. Since all the participants indicated that they have used the information on Q&A websites, we decided to conduct the content analysis of the questions which asking about the selection of the statistical tests on Q&A sites.</p>
<p>Many research analyzed different aspects of Q&A websites</p>	<p>Q&A website is often provided to corporate and specialist sites, so the site and its users can be asked questions as well as provide or receive expert answers to them. Many studies have explored the multidimensionality of Q&A websites. Most examine the content of answers within the site, the types of questions asked, and the types of users interacting [34, 1]. Some studies also focused on specific fields of Q&A websites, such as eating disorder questions [8] and music information-seeking behaviors [40], however, there is no study analyze questions about statistical information-seeking behavior.</p>
<p>Our Q&A content analysis focused on statistical test selection questions</p>	<p>In this study, we analyze statistical-related questions asking for choosing the proper statistical tests that users posted on different Q&A websites (StackOverflow, CrossValidated and ResearchGate), to explore contextual elements that characterize general data scientists' statistical information needs and information searching. Detail content analysis</p>

has been used to find out how questions are constructed, what information is provided and what problems exist in question expression.

The result indicates three motivations and six types of information provided in the questions. The missing and unclear information are the main problems of the question expression. Besides that, the information is provided in an unstructured manner, inaccurate phrase, and ambiguous sentences frequently appear in the description. These problems make the question unclear and respondents have to assume or fabricate information to answer the questions, which could lead the wrong and untrustworthy answers result shows that. Furthermore, we find that data scientists choose statistical tests in different stages of the study. This question timing is also a major factor will influence the type of information could be provided at the current stage. Based on these findings, some implications are provided for the design of statistical information searching-systems at both algorithm and interface levels.

Chapter 2

Related work

2.1 Selection of Statistical Tests

There are many previous research show issues in statistical test selection and confirm this is a challenging task for data scientists in different areas. For example, Cairns, P [12] reviewed 41 papers from 2005 and 2006 BCS HCI conferences, 12 out of 41 papers used inappropriate statistical methods, the choice of tests used was either unclear or could be called into question. Nour-Eldein H [51] also reviewed family medicine articles published between 2010 and 2014 in Suez Canal University, found that wrong analyses were recorded in more than a quarter of articles as 17/60. As a result of this uncertainty method choosing and wrongly using procedure, their contribution to knowledge must be strongly called into doubt.

Study showed HCI and medical researchers used inappropriate statistical methods

A variety of research also has focused on understanding uncertainty in statistical analysis and how analysts work with uncertain data, including defining typologies, task taxonomies, coping strategies, and creating analysis and sense-making models. For example, Nadia et al. [7] conducted a qualitative user study to understand how data workers analyze uncertain data. They also describe their various coping strategies to understand, minimize, exploit or even ignore this uncertainty. Pang et al [53] define uncertainty for the scientific visualization domain to include "statistical variations or spread, errors and differences, minimum-maximum range values, and noisy or missing data". With some overlap to Pang et al.'s taxon-

Data scientists face different types of uncertainty in statistical analysis

<p>Many literature from various domains give suggestions on statistical tests selection</p>	<p>omy, Klir et al [37] describe uncertainty as a source of deficiencies such as incompleteness or as imprecise, unreliable, vague or contradictory information. Different from the research mentioned above, we mainly find uncertainty under the decision-making procedure of data workers when choosing statistical tests.</p>
<p>Other types of resources</p>	<p>Many studies try to provide the suggestion of choosing the proper statistical tests. Barun et al. [50] proposed a five-questions scheme which will cover the hypothesis testing demands of the majority of observational as well as interventional studies. Marusteri et al [45] presented a step by step guide for biostatistics about the test selection process used to compare two or more groups for statistical differences and in order to provide to the user an easier understanding of the basic concepts necessary to fulfill this task, appropriate guidance approach is presented in a Q/A (Question/Answer) manner.</p> <p>Besides from literature, there are some other materials provide useful tutorial for data scientists for choosing statistical tests. For example, the website [26] provided by UCLA shows a table which covers several common analyses and helps users choose among them based on the number of dependent variables the nature of the independent variables. Another well-known website is provided by John H. McDonald [47], he presented a table with purpose, notes and also examples to help users decide which statistical test or descriptive statistic is appropriate for their experiment. Books [44, 61] and video tutorials [27] are also frequently used by analysts for acquiring information</p>

2.2 Q&A Websites

Q&A website is often provided to corporate and specialist sites, so the site and its users can be asked questions as well as provide or receive expert answers to them. It's frequently integrated by large and specialist corporations and tends to be implemented as a community that allows users in similar fields to discuss questions and provide answers to common and specialist questions [64].

Essentially, this definition indicates three basic functions of a social Q&A website: First, it provides an interface for

users to post their questions and answers. Second, it provides a search engine that helps people find related questions in the online community. Third, users can participate in discussions in its online community

There are three types of online Q&A services: digital reference services, “ask an expert” sites, and community question and answer sites [32, 28, 60]. “Digital reference services” refer to tools for library patrons to communicate and pose reference questions to librarians via an online system. “Ask an Expert” sites are characterized by an answerer with some type of credential in a given topic area; the interaction between questioner and answerer is not one of the peers. The final category, community question and answer sites, include those sites like Yahoo! Answers that bring together peers. For Choi et al [16], there are four types of social Q&A website: community-based, collaborative, expert-based, and social.

Many studies have explored the multidimensionality of Q&A websites. Most examine the content of answers within the site, the types of questions asked, and the types of users interacting. For example, Adamic et al’s [1] analysis of Yahoo! Answers revealed that users are seeking more than hard facts, but also advice, opinions, and thoughts on questions without one definitive answer. They grouped questions into three broad clusters: discussion forums (users “both pose and answer questions”), advice (“people both seek and provide advice and commonsense expertise”) and factual answers (“people tend to ask or reply”).

Ignatova et al [34], developed their classification of question types in Yahoo! Answers based on a sample of 755 questions about data mining, natural language processing, and e-learning, offering one of the few, if only, a taxonomy that is grounded in a specific subject domain. Their question framework includes categories such as concept completion, definition, procedural, comparison, disjunctive, verification, quantification, causal, and general information need.

Some previous works also discuss what motivates people to answer questions in Q&A websites. Choi et al [17] surveyed 75 users regarding their motivations, methods,

There are three types of Q&A sites

Different taxonomy of questions on Q&A sites

One study showed motivations of questioners

and expectations relating to asking questions within Yahoo! Answers. Results indicated that the primary motivation for social Q&A website users asking questions was to fulfill cognitive needs—a particular need that is related to strengthening information, knowledge, and understanding.

Some studies
focused on special
domain of Q&A
websites

Of particular relevance to our research are the studies focused on specific fields. For example Bowler et al [8] focused on understanding teens' use of Q&A for health information through a content analysis of eating disorder questions collected from Yahoo! Answers. Then they identified a range of needs and motivations appearing in the questions, and developed a taxonomy of question types consisting of five overarching themes. Lee et al [40] investigated cross-cultural/multilingual music information-seeking behaviors and reveals some important characteristics of these behaviors by analyzing 107 authentic music information queries from a Korean knowledge search portal Naver (knowledge) iN and 150 queries from Google Answers website. Zhang [65] explored contextual factors of consumer health information searching by analyzing health-related questions on Yahoo Answers. Particularly, they looked at the following factors: linguistic features of the questions that users formulated, their motivations for asking the questions, the time when the questions were asked, and their cognitive representations of the problem space.

No research focused
on statistical topic on
Q&A sites

However, less research has examined how data scientists ask questions on Q&A websites, how they choose appropriate statistical tests, what information they provide and also what are the motivations. Our work is unique in that we focus on the questions related to the selection of statistical methods, a professional topic that is relevant to all data scientists.

Chapter 3

Methodology

This chapter will outline the methodology for this research. The relevance and particular goodness of fit for qualitative research and grounded theory will be explored. We will introduce two methods: the interview and the online content analysis in detail separately.

interview and the online content analysis were used in the study

3.1 Motivations

Choosing the appropriate statistical methods could be a really complex thing. There are many previous research show issues in statistical test selection and confirm this is a challenging task for practitioners in different areas[12, 51]. For example, Nour-Eldein H [51] reviewed family medicine articles published between 2010 and 2014 in Suez Canal University, found that wrong analyses were recorded in more than a quarter of articles. Another research finished by Cairns [12]shows the problem of reporting or analysis that undermined the value or the validity of the statistical testing and hence the research findings in the HCI field.

Studies have indicated issues in statistical test selection

Choosing the appropriate statistical methods could be a really complex thing. People fell overwhelming when making decision for statistical test selection. A variety of research has already focused on understanding uncertainty in statistical analysis and how analysts work with uncertain data, including defining typologies, task taxonomies, coping strategies, and creating analysis and sense-making models.

Many scientists indicate that although a variety of studies and materials [50, 45] try to give suggestions on choosing the proper statistical tests, there is still no one recognized standard that everyone can use in practice. There are so many uncertainties and complexity exist in the procedure of statistical tests selection, how do practitioners deal with these difficulties and search useful information helping them make a proper decision?

Considering all the situations list above we decide to use the qualitative methodology in our study. First we talked to data practitioners to find how they choosing statistical analysis procedure, what are their concerns. Then we did Q&A websites survey to analyzed how they ask questions related to statistical tests selection, what information they provide and what are potential problems.

Qualitative approach is especially helpful in taking an in-depth examination of an issue or topic that is complex or detailed

There are several reasons why qualitative methodology is a better fit for this study. It's obvious that choosing the appropriate statistical test is a complex and uncertain thing which is a general problem widely recognized by data workers from different fields. Qualitative studies are understood to be especially helpful in taking an in-depth examination of an issue or topic that is complex or detailed [13].

While quantitative research also examines complex issues, qualitative research utilizes methods that place greater emphasis on "processes and meanings that are not experimentally examined or measured. . . in terms of quantity, amount, intensity, or frequency" [42]. Through qualitative methods, detail and complexity can be explored at great length and depth through extended open-ended interviews and content analysis.

Qualitative approach is suitable for topic that there is little existing research

A qualitative approach is also consistent with our research topic. Qualitative research is suitable for exploratory research such as this study. Exploratory research is necessary when there is little existing research on the subject matter [19]. As there is no research which focuses on how data science workers search for analysis procedure, an exploratory approach is warranted. When there is little empirical research available on a topic, it is important to engage directly with relevant research participants to gain their insight into their experience of "everyday practices and ev-

eryday knowledge referring to the issue under study”[39]. The purpose of this engagement is to develop a ‘thick description’ [41] of the research concepts and the relationship that exists between them, rather than causal theory development. In the following part, we will describe two different qualitative study procedures in detail. First one is semi-interviews, the other is online content analysis.

3.2 Interviews

The purpose of conducting interviews in qualitative research is to “understand the world from the subjects’ points of views, to unfold the meaning of peoples’ experiences, and to uncover their lived world prior to scientific explanation” [39]. Mason [46] suggests that interviews are an appropriate data collection method when the researcher believes that the primary way to understand the phenomena of interest is by interacting directly with people to “access their accounts and articulation”; when consideration of contextual and situational factors is particularly important for understanding the phenomena; and instances when the research topic “may be complex, or may not be clearly formulated in the interviewees’ minds in a way which they can simply articulate in response to a short standardized question” [57]. These three bases for the use of interviews as a data collection method are applicable to my research. Given the exploratory nature of the research, it is appropriate to engage directly with data worker participants about how they get useful information for their statistical analysis, what are their concerns and how they use this information in the decision making process of a statistical analysis; and to examine through this interaction the contextual and other factors which may affect its use in these processes. Further, as decision-making process of choosing the right statistical test is a complex topic which is not observable, nor often reflected upon or easily articulated by people given its tacit nature, it is necessary to engage directly with data scientists to explore with them their experiences of information gathering processes in order to understand how they find and use this useful information for data analysis procedure.

Directly talking with data scientist is better to understand complexity and contextual factors of statistical selection

Semi-structured interview can get in-depth information about the phenomena and subjective viewpoint of participants

Between the continuum endpoints of structured and unstructured interviews lies a multitude of research positions. However, in our paper, we explore the intermediate space of the semi-structured interview, the most common of all qualitative research methods [3]. Semi-structured interviews involve asking participants key (often predetermined) questions based around the topics of interest, and encouraging them to provide open responses to these questions [48]. This method is not as limiting as structured interviews, and consequently provides in-depth information about the phenomena of interest [10]. It is also more focused than the completely unstructured interviewing method, thus restricting the amount of potentially erroneous data provided by the participants, and enabling the researcher to collect “relevant, valuable and analytically rich data”[6]. This tighter structure assists the data coding process, and enables more effective analysis and comparison of data that would occur with completely unstructured interview methods [19]. Flick [25] argues that semi-structured interviews are a particularly appropriate data collection method when the researcher is looking to access the subjective viewpoint of the research participants.

Semi-structured interview can get inside the research world of data scientists

In our study, semi-structured interviews help develop an understanding of the ways in which data scientists looking for information and choosing the right statistical method with the help of this information. The issue becomes how to get inside the research world of data workers so that the researcher is able to interpret this research world from within [59]. Utilizing an ethnographic approach to questioning, researchers can learn about organizational culture from different individuals’ points of view thus bringing into the open an often hidden environment. Many issues, such as data workers concern or information-seeking behavior, can be studied using such an approach.

3.2.1 Participants

Purposeful sampling technique was used in recruiting participants

The sample for this interview study is data scientists who engage in data analysis activities with varying levels of skills. Participants were sampled primarily using ‘purposeful’ sampling techniques[55]. Purposeful sampling is a standard initial method of sampling in qualitative re-

search [11]. It involves the purposeful selection of “participants or sites (or documents or visual material) that will best help the researcher understand the problem and the research question” [19]. It is used particularly to identify “information-rich cases who will illuminate the questions under study” [55]. This was certainly the point of the selection of data workers for this study given their involvement in, and experience of, information seeking and decision making process with useful information.

The key criteria for the number of participants for the study are that theoretical saturation is obtained. For purposes of this study, the initial goal is for 12 participants, but is contingent on the data analysis. True to the flexible style of the approach, more or less may be appropriate depending on how and when theoretical saturation is obtained. In a dissertation study examining supervision processes in applied settings, Lonneman-Doroff [43] found eight participants sufficient for theoretical saturation, but interviewed three additional participants to assure saturation. Similarly, for dissertation research using grounded theory methods to examine how supervisees experienced the process of developing a theoretical approach to counseling, 15 participants were sufficient [46]. The participants will be recruited through purposeful sampling based on their having experience in data analysis, chosen as best being able to represent and elaborate on the topic of interest for this study [19]. We recruited participants by email contacts at organizations within our personal and professional networks.

Data researchers from HCI, psychology, statistics, and economics at different universities from different countries will be emailed asking if they are willing to participate in the study. In order to find generalized information in different areas, we also take data workers for industry field into consideration. Contact information of participants from industry is gathered through their previous classmates and teachers.

We interviewed 12 domain experts aged 24–29 from different organizations (2 enterprises, 10 research). The organizations were from different sectors including HCI, psychology, computer science, economics and statistics (Table 1). They held a number of job titles, including “researcher”, “engineer”, “student”, and “risk analyst”. Par-

Theoretical saturation is obtained after 12 interviews

Participants come from different research and enterprise domains

Participants ranged from master students in their third year of work to data analyst with over 3 years of experience. In this paper, we use the term “data worker” to refer to anyone whose primary job function includes working with data to answer questions that inform business or research decisions.

Table 3.1: Interview participant profile

ID	Age	Organization	Domain
P01	27	Research	International business statistics
P02	25	Research	Applied psychology
P03	25	Research	Business statistical analysis
P04	28	Research	Human-computer interaction
P05	25	Research	Cognitive neuroscience
P06	26	Enterprise	Computer science
P07	29	Enterprise	Risk and model analysis
P08	27	Research	Statistics
P09	27	Research	Human-computer interaction
P10	26	Research	Automotive engineering
P11	24	Research	Psychology
P12	28	Research	Labor economics

3.2.2 Data Collection

Semi-structured interviews will be used to gather data. Each participant will be sent Informed Consent forms. Participants will engage in a face-to-face or Skype interview that will last an undetermined amount of time, but estimated to be 25-50 minutes. The length of the interview will depend on the amount of data gathered. The interviews will be recorded using a hand-held recording device or a computer program, depending on the available and preferred communication medium of the participant. Given the semi-structured format of the interview, a series of pre-determined, constructed questions will be asked of participants. Additional questions will emerge at the moment as the interview proceeds, contingent on the need for clarification, in the exploration of information given, and dependent on the experience level of the participant.

We tried to understand all participants recent study including statistical analysis, the questions cover different aspects such as analysis procedure, problems they met, available resource and attitudes towards statistical analysis. Follow

Median Interview
time is 40 mins

up example questions were asked to probe or clarify responses.

- (1) How long have you been doing the statistical analysis?
- (2) What's is your recently completed study included statistical analysis?
- (3) Which test method did you choose?
- (4) How did you decide which test methods to use?
- (5) Are you sure or confident about if you choose the right test methods?
- (6) Normally how you find information which is useful for your statistical analysis?
- (7) Do you think the useful material is easy to find for statistical analysis?
- (8) Which tools did you use for this statistical analysis study?
- (9) What do you think the most difficult part of this statistical analysis?

3.2.3 Data Analysis

We analyze interview data with the framework of grounded theory. Grounded theory is an approach that is used specifically to explore relationships and complex phenomenon that has not undergone substantial consideration [15]. Additionally, grounded theory is also ideal for research questions that require an explanation of a process or an interaction, as is the case for this study [19]. Given the lack of research on this topic from the sample of interest for this study, along with the goal of gaining a thorough understanding of the information seeking and decision-making behavior of data workers, qualitative methods utilizing grounded theory is most appropriate.

Unlike other research methods, the grounded theory does not provide hypotheses for the research questions prior to gathering the data, as grounded theory is explicitly emergent and does not test a hypothesis [23]. The "emergent"

Grounded theory was used to analyze the interview data to find emergent theme

characteristic of grounded theory indicates an inductive style of research and data analysis, which allows for new properties of the process or phenomena being researched to emerge [15]. The ability to recognize novelty outside of preconception, as well as the ability to interact open-mindedly with data, maybe hindered or restricted by a narrow focus on specific hypotheses. While the method of research is not detailed in length here, it is important to note this important distinction of the methodology that precludes any statements of hypotheses at this point. It is a fundamental tenet of the approach that hypotheses not be provided in order to ensure that, as much as possible, the emergent theory is derived from the data and not from another source or preconception [20].

Data coding involves two phases: initial coding and focused coding

Data coding involves two phases: initial coding and focused coding, which cumulatively constitute the process and progression of defining the data to making meaning of the data. Initial or open coding involves labeling the data – by word, byline, or by paragraph – while focused coding uses the initial codes to “sort, synthesize, and organize large amounts of data” [14]. Labeling the data by word, line, or paragraph is indicative of the interactive relationship the researcher has with the data. Through the process of focused coding, tentative theoretical categories emerge [20].

Memo writing helps researchers remember emergent thoughts and questions

Analysis of the data is facilitated through the process of Writing memos helps the researcher to process information, to explore alternative lines of reasoning, and to more fully develop connections between categories [14]. Aside from the analytical benefit of memo writing in connecting ideas, it also reveals gaps in the data collection and helps document theory development through the research process.

MAXQDA was used to conduct qualitative coding process
Coding the qualitative data creates structure

Then the qualitative coding process was conducted with MAXQDA Standard 2018. This software organizes and manages the data entered to efficiently target the data analysis. First, I started to analyze interview transcripts with the initial coding process which is the initial step of the analysis, I carefully read the text, marked interesting and revealing selections from the interviews. I also wrote memos next to text section or passages. Highlighting tools were available with a palette of color in Maxqda (See Figure 3.1). Categories are established After the first round of

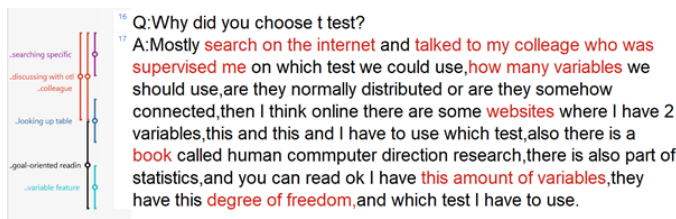


Figure 3.1: Coding process with MAXQDA 2018

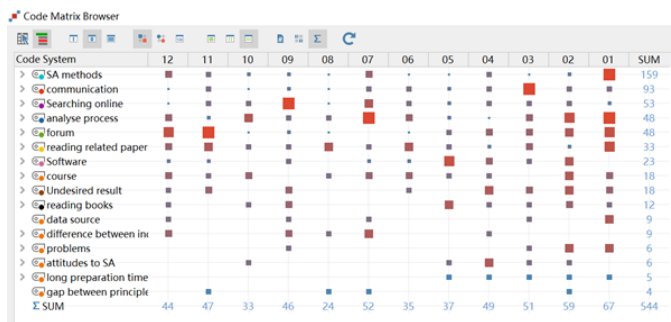


Figure 3.2: Main code matrix

the coding process. Within each category, I selected items with common characteristics that differentiate them from the main category and formulated a definition for each new subcategories. Then I started focus (second) coding process with going through all data again and assigning text to the newly created subcategories. Once codes system were established, themes were examined and questions emerged.

Finally, through theoretical sampling, all theoretical possibilities of explaining the data are considered, and through theoretical saturation, the data remains unchanged with additional exploration [15]. After the process of data coding and memo writing, we have in total of 544 codes belonging to 14 main categories which cover different aspects, such as participants analysis procedure, problems they met, available resource and attitudes towards statistical analysis. Main categories and their frequencies in each interview transcript are shown in Figure 3.2.

544 codes and 14 main categories was generated after qualitative coding

3.3 Content Analysis of Q&A Sites

3.3.1 Websites Analyzed

- **Stack Exchanged : StackOverflow¹ & CrossValidated²**

Stack Exchange is a network of Q&A websites on topics in diverse fields.

Stack Exchange is a network of 174 websites that are created and run by experts and enthusiasts who are passionate about a specific topic. The Stack Overflow and CrossValidated are two websites run by Stack Exchange. Stack Overflow is an online platform where users can exchange knowledge related to programming and software engineering tasks. While as a searchable Q&A site oriented toward statistical analysis, CrossValidated is a question and answer site for people interested in statistics, machine learning, data analysis, data mining, and data visualization.

Tags helps us filter interested topics

Stack Exchange features the ability for users to ask new questions and answer existing questions, as well as to “vote” questions and answers up or down, based on the perceived value of the post. Users can earn reputation points and “badges” through various activities. All questions are tagged with their subject areas, users can click any tag to see a list of questions with that tag, or go to the tag list to browse for interesting topics.

Edit function helps us figure out improper information

It also provides edit function which users can fix mistakes, improve formatting, or clarify the meaning of a post and comment function which users can ask for more information or clarify a question or answer. These two functions are really helpful for our analysis, since we can analyze what is improper description of the question with edit function and what is missing and unclear information in comments.

- **ResearchGate³**

With fifteen million academic users as of 2019, ResearchGate is one of the most well-known academic social networks. Its Q&A model is where researchers

¹StackOverflow: <https://StackOverflow.com/>

²Cross Validated: <https://stats.stackexchange.com/>

³ResearchGate: <https://www.researchgate.net/>

can ask research-related questions and get them answered by other specialists. It's also the best place to share the knowledge, connect with researchers.

Many academic practitioners indicate that it is hard to find statistical experts in real life, and ResearchGate helps them solve this problem. ResearchGate's Q&A recommendations take researcher's unique set of skills and expertise into account to present users with the most relevant discussions in your field. Researchers can browse questions and answers using the three filters: recent questions in the specific field, questions followed, and questions asked.

ResearchGate is widely used by academic users to ask questions and get domain expert's answers

3.3.2 Data Collection

Our research questions involve analyzing user-generated posts about statistical tests selection on StackOverflow, ResearchGate, and CrossValidated. We also decided that our method should be manual content analysis because the types of analyses we want to perform on the questions and responses cannot be obtained through either quantitative methods or automatic methods.

Because information behaviors on statistical tests selection Q&A are unknown in the literature, we decided to start with one general discipline for developing our analysis scheme, then expand to other disciplines. Since the purpose of this analysis is to investigate how data scientists seek information about selecting proper statistical test, our first discipline is the selected question should focus on procedure of choosing appropriate statistical method.

Selected questions should focused on procedure but not implementation

Based on this selection criteria, we first use keyword 'statistical test' to search meeting standard posts on both StackOverflow and CrossValidate. In the filtering process, we found that it is difficult to find out questions' description problems and judge the integrity of the provided information only by analyzing the questions. So the second principle was added: The post should have at least one reply or comment from the respondents. In order to make the study general, different keywords from a variety domains

'statistical test', 'significant test', and 'hypothesis testing' are three filtering keywords

were used to filter the questions: 'statistical test', 'significant test', and 'hypothesis testing'.

At last 101 questions across different domains and platforms were selected and Google sheet⁴ is used to record links and detail finding information of each question, so that every question and information are traceable.

3.3.3 Data Analysis

Missing/unclear questions was judged by reading the respondent comments/answers

The qualitative content analysis was used to analyze the content of the posts. The content analysis aimed to investigate the characteristics of questions in relation to the dual aspects of providing information. We tried to figure out what information is provided by data scientists and what are missing and unclear information that make respondents confused and cannot give proper suggestion answers. It is necessary to note that the criteria of judging missing or unclear questions is based on respondents request and expression,so we choose posts which have at least one comment or answer.

Both questions and comments/answers were coded to find emerging themes

First we read questioning messages several times to gain an overview of the overall content. Then codes were derived inductively by closely reviewing themes appeared in each message. Whenever a contextual description appears, such as the motivations of asking the question, attitude involved, and difficulties in expressing their needs, it was coded into a category. When coding a new text to a category, the text was compared to those already assigned to that category [29]. The constant comparison method allowed us to fully understand the theoretical properties of the category. We then examined the categories resulted from the open coding process to make sense of the properties and dimensions of the categories, identify relationships between them, and uncover patterns [9].

It is important to mention that the after analyzing all 25 StackOverflow questions,we found that all these

⁴All QA analysis result included in this link: https://docs.google.com/spreadsheets/d/1xSt12wIbXNa3kFIqjE55_gNxO4T-ri15GS49hgTwhYE/edit?usp=sharing

questions are related to implementation asking for programming procedure with code scripts, which violates our original intention of research. Although we analyzed these questions, at final summary stage we decide not to include them, so that in the following finding and discussion chapters, only 76 questions from ResearchGate and CrossValidated are included and discussed.

101 questions were
in detailed
analyzed,76
questions were used
to form findings

Chapter 4

Findings

4.1 Interview Findings

The purpose of this interview study is to understand how data scientists find useful information in order to choose the statistical analysis method. The analysis of the interview transcripts reveals findings structured around four main aspects: Procedure, problems, resources, and attitudes.

In this section, findings from participants' responses to the semi-structured interview questions will be presented. Analysis results in the emergence of the following themes: (1) problems in methods selection, (2) coping strategies, (3) resource utilization. Each main theme has several key findings and their interpretations emerge from coding the interview transcripts. Findings are further supported with quotations from interviews with participants.

Interview findings centers around four main themes

4.1.1 Theme 1: Problems in Methods Selection

The first theme centers around the general problems that were identified by analyzing participants' interview transcripts. In this section, two key findings from this theme are discussed.

“There is no standard!”—Uncertainty with test selection.

Data scientists need assurance with the statistical analysis

Selecting a significance test involves a complex myriad of decisions

they perform since using an incorrect method can render their findings invalid. In real-world, this might result in a publication getting rejected or an analytics firm's credibility going downhill. In our analysis, we find that participants have a sense of uncertainty with the selection of statistical analysis methods. First and foremost, selecting a significance test is not an easy process—it involves a complex myriad of decisions that lead to one or more “valid” tests that can be performed. Data is often not as ideal as we expected, sometimes these situations could happen: what if some independent variables are normally distributed, but others are not? How to handle an unbalanced number of participants in the study? What if there are outlier data in your study, will you keep them or delete them?

Data scientists feel uncertain

Given how difficult this can be, it is not surprising that data workers feel overwhelmed and uncertain:

“There is a lot of uncertainty with test selection and there are no standards. For me, it is hard to know which test to use.” - P04.

Interpreting the results thus becomes an issue

As a further complication, even when data workers do eventually manage to do a significance test, because they are uncertain about the test, interpreting the results becomes an issue as well:

“Because there is so much uncertainty about the selected test, I am not sure about the result. The whole process is complex and unclear to me.” - P11

Uncertainty causes anxiety

As we can see, all this uncertainty can lead to anxiety in data workers. In some situations, this can be so extreme that data workers have completely given up all hope and postpone the selection procedure:

“I never start by selecting the significance test to analyze my study data, because I'm not sure I can do it properly!” - P01

Multiple resource can not relieve anxiety

One might wonder whether this uncertainty can be alleviated by getting informed—there are plenty of useful resources[26, 44, 50, 61] out there that help data workers select the correct statistical test. However, finding the “right” information from these resources, especially books, takes time and, even then, there might be conflicting information

in these resources. As a result, data workers are still uncertain:

“Even if I read all these things, I would still be not sure...especially if it is just me making the decision.” - P09

This leads to an interesting question—are data workers more confident in the significance test selection if another researcher, particularly one who’s experienced in statistics, approves of it? We will discuss this in the following subsection.

Using/recommending wrong methods to others.

In addition to the sense of uncertainty, using wrong methods is another general problem indicated by some participants. Practice of wrong or inappropriate statistical method is a common phenomenon in the published articles in biomedical research. Incorrect statistical methods can be seen in many conditions like use of unpaired t-test on paired data or use of parametric test for the data which does not follow the normal distribution, etc. One participant mentioned that:

“There is another example, we analyze factors that affect the winning percentage of each team in the NBA, I just choose the wrong method for it.” - P03

Another researcher who is also a supervisor in his institute said he has not used the wrong test for his own situation but has recommended a wrong test for others:

“I have recommended the wrong method to another person.” - P04

Using wrong methods

Recommending wrong methods

4.1.2 Theme 2: Coping Strategies

The second theme revolves around the standard for selecting statistical methods. In this section, we demonstrate and explain four findings regarding what are the subjective influencing factors when people make choices. We do not focus on the objective factors which influence deciding the right test to use, such as study designs, the main study hypothesis, type of data, and independence of variables. We

Psychological activities rather than objective factors

want to analyze subjective factors, such psychological activities hide behind the surface, but have a huge influence on making the decision of statistical tests.

Choosing the most commonly used method in their research field.

Statistical tests are generally categorized into various types depending upon the type of field. Statistical tests are carried out extensively in HCI, psychology, and medicine. Ten out of twelve participants demonstrated that they prefer using the most common method in their research or work field.

each domain has its
commonly used
method

In the field of psychology, HCI and medicine, statistical tests of significances like t-test, z test, f test, chi-square test, etc., are carried out to test the significance between the observed samples and the hypothetical or expected samples. Two Ph.D. participants who are conducting research in psychology and HCI fields indicated that:

statistical tests in
HCI

“Generally there are fixed routines when you doing such projects, normally we will use ANOVA because it’s easy and better to use and it’s a common method in psychology experiments.” - P05.

“Second criteria is that the test is frequently used in the HCI field. I mean like chi-square test and t-test are all tests commonly used.” - P09.

statistical tests in
computer science

In the field of computer science such as machine learning, more complex statistical models always used to frame predictive modeling problem and better understand the data. It would be fair to say that statistical methods are required to effectively work through a machine learning predictive modeling project. Some statistical methods can be used to clean and prepare data ready for modeling. Others statistical hypothesis tests and estimation statistics can aid in model selection and in presenting the skill and predictions from final models. One master student who is learning computer science mentioned in the interview:

“There are lots of common methods in data science, such as Naive Bayes, Random Forest, Boosting, neural network, each model has its pros and cons, and

each method has its own context for use. I will always try these methods first.” - p06.

Another researcher who is a Ph.D.candidate from Labor Economics field also indicated that in his research field, people prefer using different kinds of regression models for data analysis purpose so that he will always put these methods to the first priority:

statistical tests in
Economics

“Regression is the mostly used and basic method for influence factor analysis in my research field...regression discontinuity, DID, panel data, actually they are all regression, the difference is running different codes, but the essential part is all regression.” - P12.

This also happens in the industry field. One participant working on risk and model analysis in a bank said that when he entered the company he tried with other models, but later changed to regression model and decision tree, since these two are the most commonly used methods in p2p bank field:

statistical tests in
P2P industry

“My boss told me, in p2p and bank field, the most frequently used is this model, The reason is easy interpretability.” - P07.

“There is no additive benefit.”—Negative attitudes towards learning a new method.

In addition to preferring using common methods in their field, half of the participants hold negative attitudes of learning new statistical methods. This discovery once again verified the previous finding. As mentioned before, lots of participants are not expert in statistical analysis, they just have basic theory courses at school which will not give them enough confidence for using new methods for their study. Although feeling interested in new methods emerging in the research field, one participant still demonstrated her concern in using it:

Lack confident in
using new method

“I would be interested in reading about it, but I don’t have confidence in using it, I would worry that I did something wrong because nobody used it before... I don’t feel that I’m expert enough to decide that this

Learning new method takes time	<p>is better than all others." - P09.</p> <p>Using the new method means taking the time to collect relevant information about the method, understanding and learning it so that it can be applied to their own study. This information gathering and learning procedure need adequate time period which makes people hesitate about the use of new methods. This happens in both industry and research fields:</p> <p>"If I have more rich time I will try more methods and learn some new methods. If I have time, I would like to try and learn something new." - P01</p> <p>"Because the time limitation won't let you try new things, you have lots of things to do, only the result make sense." - P07</p> <p>"If I need a long time to study it, I will reconsider it." - P10</p>
Weighing the pros and cons	<p>Before making the decision of choosing a new methodology. Lot's of people will ask this question in their mind: Is that really necessary for learning it, what are the benefits of using it? Measuring how much the new method will bring to them directly affects their decision. Two participants revealed this balance in their minds:</p> <p>"At that time they won't know what is the advantage of this approach, they have to spend a lot of time in reading the new approach, they should know the trade-off at the end." - P04</p> <p>"It unnecessary to learn new things." - P11</p> <p><i>Prefer Most familiar/confident methods.</i></p>
Some times several methods are suitable	<p>As can be seen from the previous three findings, people are conservative about new methods and like to use methods that are common in their field. The following finding better confirms and explains this phenomenon. Many times, there is more than one correct method and there may be many ways to apply to your current data and experiment design. It seems that lots of people stick to use their most familiar</p>

methods, this repeat procedure makes them feel confident about the result.

When asked about how to make a decision if there are multiple correct methods suitable for the current scenario, one researcher in HCI gave the following answer:

“There are always different approaches to do the same thing, several methods could be correct, each has their advantages, in those situations, I just do the one I’m familiar with.” - P04

Another participant who is learning business statistical analysis also provided that she prefers the familiar model than another method, but at the same time, the persuasiveness of the model is also a factor of consideration:

“I will use regression because I think regression is the one I’m familiar with and it’s also more clear and persuasive.” - P03

Compared with the two participants listed above, the following two persons didn’t hold strong attitude towards choosing the familiar methods. They want to try other statistical methods which they maybe not so familiar with, but they still put them in the first priority of the waiting list:

“I will first start with the one I’m most familiar with, it’s ANOVA, it’s really helpful and practical.” - P05

“I will try from the one I’m most familiar with, compare them and see the result” - P10

“Because the traditional method and the method which I’m familiar can solve my case. It’s more secure.” - P11

Modle’s
persuasiveness

Starting from the
familiar one and try
others

Feeling secure

Knowing there is a more advanced method before.

What makes me feel interested is that some of the participants showed that they knew there would be a better method to use which has a stronger proof of their study and explore better the relation of the data to the underlying population.

Accuracy is not
always first priority

It seems accuracy always do not have first priority, there are lots of other factors to consider when choosing a statistical method. One participant from p2p bank industry provided that:

“Actually in statistics filed, we have lots of better models, but not all the models can be used in the industry.” - P07

When he just onboard to the company, he was confused about why everyone used the regression model which is not the best one in distinguishing. Then he asked his boss who told him later that the reason is easy interpretability.

Model's
interpretability

“Actually,in industry, the most important factor is the interpretability of the model, a better result is the second factor.” - P07.

Another person also talked about this phenomenon in the interview:

“I have asked one credit data analysis company in Sweden, I asked if they use deep learning model for analysis, they said no, we don't use that complex method, we just use simple logistic regression even though former one could get a better result.” - P08.

Model's complexity

It is not a common thing which just happens in the industry, people in the research field sometimes also abandon better model which may have a better result though. One participant who has chosen the statistical method for his thesis related to machine learning illustrated in the interview:

“I used the neural network in my paper, but I also tried the recurrent neural network before, it may get a better result, but I didn't use it” - P10.

After weighing the pros and cons, he thought choosing a method he knows better is more profitable and secure so he abandoned the advanced model: recurrent neural network and use the neural network.

4.1.3 Theme 3: Recourse Utilization

Resources play an important role in statistical analysis, especially in searching for analysis procedure. Since our re-

search focuses on understanding the procedure of how data workers look and gather information, so that in this part we will discuss different resources and materials participants mentioned in the interview.

Book

There are lots of available books that cover different aspects of statistics. Such as *Statistics* [21], written by David Freedman and Robert Pisani teaches anyone the subject from the very ground up and good for beginners to get started with. Other books also introduce different applications and tests in a variety of fields. For example, *Statistics for business & economics* [4] delivers sound statistical methodology and meaningful applications that clearly demonstrate how statistical information informs decisions in actual business practice with real business examples. *Modern Statistical Methods for HCI* [58] reflects on current statistical methods used in Human-Computer Interaction (HCI) and introduces a number of novel methods to the reader.

Statistical books for specific domains

Many people mentioned they will read these books to find useful information when they choose the right statistical test. Some participants mentioned that reading statistical books is always good for systematic learning and knowing the overall methodology:

Good for overall learning a method

“If I want to learn it systematically, I will buy a book, normally is an ebook.” - P02

“I read some books about the methodology I’m going to use, I know the overall methodology, but some specifics I sort of identify on the fly.” - P04

Some other participants also provided negative opinions about reading related books for choosing a statistical test. Books always focus on traditional tests but not a good place to now up to date modern methods:

“On the book, you just stick to the old method and use them again.” - P09

Furthermore, books always contain numerous information, which will cost a huge amount of time to find the information people really need. For those who face time pressure with their study, this slow reading procedure will not be a

Negative attitudes of reading books

good choice:

“I don’t read books, I think reading is slow.” - P10

For handling this, one participant also mentioned the solution:

Goal-oriented
reading

“When I buy a book I will go through the content first, and the content is really goal-oriented, I want to do what and I find it’s included in this book so that I will buy it.”-P05

Paper

For researchers, reading the related paper in their research area is another way to get useful information. Almost all the researchers who attend the interview mentioned that they will go through related papers with a similar topic and refer to what statistical test they use for analysis, as we can see in these following extracts:

“When I got my study topic, I will read lots of reference paper, what methods they use, I will probably use the same one.” - P05

Referring methods in
papers with a similar
research topic

“I go to google scholar for searching the paper with a similar topic with my study.” - P09

“That’s the usual way. I will refer to the statistical method that literature uses. See how they analyze this kind of data.” - P11

Contrary to the previous finding which reveals books is better for learning traditional statistical methods, some people mentioned literature always a good place to know up to date methods recently emerged in your research field:

Papers are better to
know up to date
methods

“When I want to learn a new method, I mostly will find some paper which uses this method.” - P02

Two researchers also provide another application, they use similar paper to filter statistical variables for their study. One said:

"I also read some paper which has a similar topic with mine for finding the proper variables."-P10

The other also mentioned she uses paper to refer to what factor others choose for analysis:

"I will search for some paper related to this field, find the factor mentioned in the paper and for each factor what are the independent variables they used." - P03

At the same time, not all the related papers can be chosen, there are still some choosing criteria mentioned by the participants. First is publish time, they prefer to give first priority to literature published most recently:

publish time:recent
years

"I will normally consider the paper after 2005." - P03

The second factor is the journal of publishing, they think articles published on top academic journals are more credible and valuable:

publish journal:top
academic journals

"First I will see the level of literature published in journals, I will firstly consider literature on the top journal, level of literature if the level is same, I will try them to see which result is significant." - P12.

Communication

During the interview, we found that communication plays a significant role in statistical analysis for both research and industry participants. Asking and discussing with others seems to be the best way to solve statistical problems. As can be seen from the previous findings, people are uncertain about choosing the right statistical method and this uncertainty throughout the whole analysis process. The common way provided by most of the participants is discussing or asking others to validate their thoughts or get some clues in the right direction, as can be seen from the following extracts:

Discussing or asking
others to validate
their thoughts

"I will ask the senior who has used this before, if he used before, the information you got will really helpful." - P02

People used the
methods before

"I have this and this tests, and he also as well

checked again, and finally, we agreed on which test to use." - P09

"Communication especially asking others is really important in the work, normally if you find something by yourself is time-consuming, you should combine searching yourself and asking people together." - P07

Except for solving problems in statistical analysis, through communications people can get other's recommendation of the powerful method, useful resources and field's experts. For example, one industry people said to me he knew some better method from his colleague:

Knowing better
methods

"At First, I did the model myself. After communicating with my colleagues, I find some other data filtering methods." - P07

Another student mentioned that his teacher recommended some expert to him for solving his problems:

Recommending
experts

"Always I will get the answer from my supervisor if she doesn't know she will recommend me some expert person." - P11

There are different persons people would ask and discuss with. Most students said they will discuss with their mentor when choosing the statistical method, such as:

Supervisor

"I also discussed with my supervisor to see which one is more reasonable. Since my supervisor is experienced in this field." - P10

While most industry people prefer discussing with their colleagues, they think it's the most effective way to solve their current concern:

Colleagues

"Some colleagues have already worked for 10 more years in this field, some of their experience is really helpful." - P07

Hard finding experts

Some researcher also mentioned that they want to discuss with experts in a specific field and method, but they don't know these people and also don't know how to find and contact them. Hard finding expert seems to be another common issue in statistical analysis. Five out of twelve peo-

ple provided their problems of finding the expert, in the following are some extracts:

“I went to some publications such as last year CHI conference to see which paper use the same methodology and know the authors and email them. That’s again the problem that finds the right expert in the methodology.”-P04

“First I will ask my leader who is the experts in this method, my leader will recommend the person to me.” - P08

Q&A Websites

The emergence of social media technology has led to the creation of many online statistical websites, where data workers can easily share and look for statistical information from experts in specific statistical fields. For example, StackExchange-CrossValidated¹ is a Q&A site for people interested in statistics, machine learning, data analysis, data mining, and data visualization. Another famous Q&A website is ResearchGate² which is a social networking site for scientists and researcher to share papers, ask and answer questions, and find collaborators. Some others like Stackoverflow³ and R Community⁴ are also commonly used question-answering sites for data workers.

The interesting thing is that during the interview, all the participants have mentioned they have directly or indirectly used online Q&A forums to search for useful statistical information. Directly means they know the forum before and directly go to the page. In contrast, indirectly means they are lead by other sources, such as they google something and then find answer link to the page. Participants mentioned different use scenarios of these websites. Some of them ask their questions on the sites while others prefer going through similar questions. When asked about how they think about these Q&A forums, participants provided different attitudes and opinion. Furthermore, they

All participants have used Q&A sites information

¹Cross Validated: <https://stats.stackexchange.com/>

²ResearchGate: <https://www.researchgate.net/>

³StackOverflow: <https://StackOverflow.com/>

⁴R Community: <https://community.rstudio.com/>

also mentioned their criteria for judging the credibility of the answers.

All the above information from the interviews is summarized in the table 4.1.

Table 4.1: Information of Q&A Websites summarized from interviews

Category	Subcategory	Interview Extract
Using scenarios	asking questions	"I asked the question I mentioned before, related to principal component analysis. There are 4 or 5 replies within 2 days."-P12
	searching similar questions	"I will just go and read.I would see a similar question to my problem."-P09
Judging criteria	upvote number	"I will also refer to like number of the answer." - P02
	comments	"I will also see the comments if lots of people agree with it." - P09
	explanation	"But if the answer is really in detail with the explanation and theory, I will go deep into it." - P03
Attitudes	positive	"Lots of people discuss below the answer, I know lots of methods there and also some analysis mentality." - P06
	negative	"You cannot find the exact same question, I just find similar ones, compare and integrate." - P05

Since all participants mentioned they have used Question-answering forum to search for some useful information, it seems to be the most common resource used by data work-

ers. Combined with the previous findings which reveal the uncertainty in choosing the right statistical method and hard finding experts in specific statistical fields, we then did a survey with three commonly used Q&A websites to analyze more of how people provide information and ask for information, also how respondents answer their questions. Detail information will be shown and discussed in the following part.

4.2 Findings: Content Analysis of Q&A Websites

The context within which a data scientist seeks statistical information is multidimensional. We structure the findings from our analysis into four theme questions:(1)what type of questions do data scientists ask? (2)what information do data scientists provide in the questions? (3)what problems do data scientists have when providing information? (4)when data scientists choose statistical tests?

Analysis results centers around four main themes

4.2.1 Type of Questions

Unlike prior research on social Q&A websites [31, 34, 32], our framework is not meant to be a universal categorization of all questions in Q&A websites. Rather, these categories are specific to the questions asked by data scientists, in the domain of statistical test selection.

From our analysis, we identify that data scientists use Q&A forums for one of three purposes: Seek factual information, seek validation, and seek resource.

Three motivations of asking for statistical test selections information

Many questions aim to seek all of the above. This reflects that questions in ResearchGate and CrossValidated can be long. The questioner needs to write verbose texts that embed several queries and provides adequate information to the potential respondents. This may also be a reflection of the complexity of the problems associated with the selection of the statistical test.

Questions that seek factual information (53/76)

Most of the questions related to statistical tests posted on

“Look-up” question expecting quick, fact-based answer

the Q&A websites are seeking factual information. The factual question can be seen as a “look-up” question, where the expected outcome is a quick, fact-based answer. These are questions that seek to fill the questioner’s knowledge gap.

As discussed in section 4.1., it is universally acknowledged that selecting a significance test is not an easy process—it involves a complex myriad of decisions throughout the whole analysis process, from study design to reporting of the statistical result. So it is not surprising that data scientists use Q&A websites to ask questions to improve their decision-making skills.

Several queries
embedded to one
questions

Although we focus on questions that seek information on choosing the appropriate statistical test, several queries of other factual information always embedded into one question which covers different aspects and stage of the study process. For example, there are two questions related to information of sample procedure which is the stage before the statistics(‘C’ and ‘R’ mean the question number in result sheet⁵):

“[...]Company A wants to sample as few students as possible because accessing old data is time intensive. What sampling procedure should the company use?
” - C27

Two questions asking
for sampling
procedure

“[...]Total population is 57, How to identify sample size and which sampling technique to use?”-R7

Just understanding proper test for the specific circumstance is always not enough, for those who just step in the statistical field, following implementation procedure is also a knotty problem. Here is one question requiring implementation information:

One question asking
for implementation of
the test

“My data consists of [...]Which statistical test should I apply here? I was thinking repeated measures ANOVA, but am not sure, and don’t know exactly how to implement this in R.”-C8

⁵All QA analysis result included in this link:https://docs.google.com/spreadsheets/d/1xSt12wIbXNa3kFIqjE55_gNxO4T-ri15GS49hgTwhYE/edit?usp=sharing

For researchers, using the statistical result to validate their finding is an important skill. Good interpretation of the data always adds points to the final report. One asker asks for factual information about how to interpret the result in the community:

“[...]What statistical method is fit for such circumstance and how could we possibly interpret the result of the statistical treatment?”-C6

One question asking for result interpretation

Seeking Validation (36/76)

Nearly half of the questions related to statistical tests posted on the Q&A websites are seeking validation from experts. This phenomenon is verified and in line with our previous interview finding. As discussed in section 4.1 theme one, selecting proper statistical tests is always a difficult and complex process, data workers show overwhelmed and uncertain in the interview transcript.

In questions where validation is sought, questioners have already analyzed the situation and read material to build initial cognition, then they present an experiment procedure, an opinion of the proper statistical test, or explains how they initially process their data and seeks confirmation from the answering websites.

Questioners already have one or more choices

In some cases, questioners met problem after collecting the data, they may find out that the data does not met their expectation which will cause an uncertain of previous experiment procedure or initial data analyzing method. For example, one questioner explained detail information of the experimental design to describe how the data collected, but after finishing the initial data analysis step, confusion occurred.

Special situation happens to the study

In other cases, many questioners already make the decision in mind, but some specific situations such as small sample size or unconfirmed data type lead them to ask for validation from experts. One asker wanted to know if it's valid to do a mixed ANOVA when the between-subjects grouping factor has very unequal group size:

“[...] Is there a cut-off point where group sizes are so unequal that there is no validity (or point) in doing statistical comparisons (ANOVA in particular) or is

there a way to analyze this 2x2 design despite the large group sample differences?"-R5

Result is not in accordance with expectation

As a further complication, since data workers are not confident about choosing the proper test, if the result is not in accordance with expectation, suspecting the test and interpreting the results becomes an issue as well:

"[...] Based on previously published data a one way ANOVA would be best for what I am doing...When I include steroid A on the graph nothing is significantly different from my basal reading...This seems incorrect to me as this means there is dependence between each response...Is this because of the incorrect test and which test should I then use? Or if this is correct, why should the bar for steroid A be effecting the statistical analysis of my remaining steroids?-R10

As can be seen from the above examples, choosing appropriate statistical tests is really complex and need data workers to make multiple decisions. Even the experienced data worker could face uncertainty in this process and search for confirmation from others on the websites.

Seeking Resource(6/76)

One question asking for tutorial with example

A small part of the questions on the two Q&A websites requires resource for building the cognition of the statistical test. As discussed in section 4.1 Theme three, data workers reveal in the interviews that they use a variety of resources in statistical analysis, especially in searching for analysis procedure. Compared with the previous finding, new resource types emerges in the Q&A websites. There are two types of resource questioners require in our analysis. First is tutorial or material with examples. Four questioners ask for reference examples, which validate our interview findings that data scientists prefer using examples to learn an unfamiliar method. One example is:

"[...]you are mentioning the significance of the whole model. Is there a particular procedure you would follow? Or do you have a link/example you could share?"-C32

The other one is code scripts or package. Knowing what statistical test to use is just the first step, how to implement the test with software is also important for data workers. In our analysis, one asked for specific code script in the website:

Code scripts

“[...]Can we do any type of statistic to see in such case? if yes please give me code with respect to R.”-R21

In another question, the questioner wanted to know which R package can be used to implement the test:

One question asking for programming package

“[...]Which model would be suitable for my problem (probably a multilevel linear model?) and which R packages I could use for the analysis?”-C12

It's necessary to note that this requirement not just put forward in original questions, but also in the comments after one answer giving the suggestion test.

4.2.2 Information in Questions

Second theme centers around the type of statistical information occurred in two Q&A websites of our surveys. A summary of what information data workers have provided in their questions will be represented. Then the overlooked information which was required by the respondents afterward will also be discussed.

Matrix of the provided information

The information that askers provided in the questions is not only a representation of their understanding of the current conditions but also reflects problems of choosing the proper statistical methods. Table 4.2 shows the types of information that askers provided in their questions. The frequency is the number of posts that contain each type of information.

As can be seen from the table, data scientists mainly provide six types of information to present their problems in choosing statistical tests.

Data practitioners provide six main types of information

Table 4.2: Types of information that askers provided in the questions

Provided Information		Frequency	
Dataset (32/76)	All raw data	13	
	Part of raw data	7	
	Fabricated data	3	
	Summary data	11	
	Assumption	14	
Experiment Design (76/76)	IV (71/76)	Name	66
		Other detail	66
	DV (64/76)	Fabricated	6
		Name	64
		Other detail	19
	Sample size		32
	Procedure		17
	Within/Between-group design		19
Hypothesis		26	
Goal of analysis		67	
Software		17	
possible solution		36	

Dataset

Summary data are used to summarize a set of observations

Many questioners also provide summary statistics to describe their datasets. Summary data are used to summarize a set of observations, in order to communicate the largest amount of information as simply as possible [63]. Summary statistics could be a measure of statistical dispersion, the shape of the distribution or statistical dependence. Table 4.3 shows different types of summary statistics emerged in our analysis.

Table 4.3: Types of summary statistics that askers provided in the questions

Type of Summary Statistics	Sample Questions
A measure of statistical dispersion (eg.mean and standard deviation)	"[...]Additional information: -The mean of A: 7.11dollars and B:7.49 dollars -The standard deviation of both is the same \$8.00"-C34
A measure of the shape of the distribution (eg.skewness or kurtosis)	"[...]The skewness and Kurtosis are under acceptable range but the Kolmogorov-Smirnov normality test shows significant value"-R27
If $n > 1$ summary statistics could also be a measure of statistical dependence (eg.interaction or correlation coefficient)	"[...]When I ran the ANOVA with the between-subjects factor however,the within-subjecteffects came out not significant and so did the interaction between duration and handedness."-C3

Many data analysts also provide assumptions of the dataset. Usually, in inferential statistics, certain assumptions need to be assessed prior to analysis. Many statistical tests have assumptions that must be met in order to ensure that the data collected is appropriate for the types of analyses you want to conduct. Depending on the statistical analysis, the assumptions may differ. A few of the most common assumptions in statistics are normality, linearity, and equality of variance. Failure to meet these assumptions, among others, can result in inaccurate results, which is problematic for many reasons. When testing hypotheses, running analyses on data that has violated the assumptions of the statistical test can result in both false negatives and false positives, depending on the particular assumption violated. Table 4.4 shows three common types of assumptions and their sample questions in our analysis.

Many statistical tests have assumptions that must be met

normality, linearity, and equality of variance are the three most common assumptions in statistics

Table 4.4: Types of assumptions that askers provided in the questions

Type of Assumptions	Sample Questions
Normality	"[...]I have 3 independent variables, they are not normally distributed and they violate few assumptions such as the linearity violations of independence."-R24
Linearity	"[...]I do a regression over the data, but it seems to be non-linear. Which test do I have to do now to test my hypothesis?"-C31
Equality of variance (Homogeneity of variance)	"[...] I can't assume that the variance is the same,so that cancels out the paired t-test and the homoscedastic t-test."-C10

Many statistical tests have assumptions that must be met

normality, linearity, and equality of variance are the three most common assumptions in statistics

Many data analysts also provide assumptions of the dataset. Usually, in inferential statistics, certain assumptions need to be assessed prior to analysis. Many statistical tests have assumptions that must be met in order to ensure that the data collected is appropriate for the types of analyses you want to conduct. Depending on the statistical analysis, the assumptions may differ. A few of the most common assumptions in statistics are normality, linearity, and equality of variance. Failure to meet these assumptions, among others, can result in inaccurate results, which is problematic for many reasons. When testing hypotheses, running analyses on data that has violated the assumptions of the statistical test can result in both false negatives and false positives, depending on the particular assumption violated. Table 4.4 shows three common types of assumptions and their sample questions in our analysis.

Table 4.5: Types of assumptions that askers provided in the questions

Type of Assumptions	Sample Questions
Normality	"[...]I have 3 independent variables, they are not normally distributed and they violate few assumptions such as the linearity violations of independence."-R24
Linearity	"[...]I do a regression over the data, but it seems to be non-linear. Which test do I have to do now to test my hypothesis?"-C31
Equality of variance (Homogeneity of variance)	"[...] I can't assume that the variance is the same,so that cancels out the paired t-test and the homoscedastic t-test."-C10

It's necessary to note that among these thirty-two questions with datasets, three of them were provided after requesting by the respondents in the comments.

For the rest questions which don't have any type of dataset provided, more detail analysis then conducted to find whether the provided information is enough for respondents to give credible answers and what potential problems exist behind this questions. Table 4.5 summaries the following scenarios and findings of those questions without datasets.

Table 4.6: Question without dataset

Information type	Scenarios	Sample Questions
Enough for respondents (34/44)	Provide enough info about the variables or data collected procedure. (27/34)	"[...]I am trying to study the association between blood groups (categorical variable) and cholesterol levels (continuous variable) adjusted confounders? I was wondering which statistical test would be more appropriate?" -C5
	Questions focus on solving problems which don't need datasets. (7/34)	"[...]Is it meaningful to test for normality when my sample size is so small, or can I simply assume normal distribution?" - R15
Not enough for respondents (10/44)	Respondents request the dataset (2/10)	"[...]Can't respond responsibly without more information. Please show some data, or speculations of how data may look." - C15
	Respondents assume the dataset (8/10)	"[...]I suspect the distribution of hospital costs is positively skewed, is it? If so, I would strongly consider using quantile regression rather than OLS linear regression." - R6

Experiment Design

No matter how concrete it could be, all of the questioners give information about experiment design.

The experimental design is the branch of statistics that deals with the design and analysis of experiments. The designing of the experiment and the analysis of obtained data are inseparable.

The designing of the experiment and the analysis of obtained data are inseparable.

In an experimental study, variables of interest are identified. Most of the questioners prefer providing information of independent variables and dependent variables, because variables information is important in choosing statistical tests. The UCLA faculty [26] summarized a table which shows general guidelines for choosing a statistical analysis based on numbers and nature of both independent and dependent variables.

Most of questions provide variables information

The result of our analysis reveals that most of the variables were provided with the real name of the study scenarios. However, still some of the questioners fabricated the variables when asking the questions. There are three types of fabricated variables information emerged in our analysis, detail information will be discussed in the following part.

There are three types of fabricated variables information

In some questions, askers created example variables with similar nature to their own variables, but they were not willing to provide the real information, here is one example:

Fake variables with same properties

“Suppose I am doing a case control study. Lets say...For Example: If “satisfaction with life” and “quality of life” are research variables in two groups...Please explain as my research question is different, and I have just given an example here.” - R11

In some other cases, although the real variables name was provided, the data scientists still fabricated the nature of variables. Such as in the following question, the questioner gave example categorical dependent variable with three categories, but later commented that the real dependent variable had five categories:

Fake properties

“I should say that my data has two more score values than in my example above, it was designed exemplary” - C1

The last fabricated scenario is that some questions just provide general information of the variables, but without real name and context of the study. In one example question

General description without real name and context

showing below, we can just infer from the description that there are two variables with Likert items, but we don't know what these variables are and the context background information:

"I am running Kruskal-Wallis tests in SPSS to compare answers to Likert items among 3 (variable 1) and 4 (variable 2) groups." - C29

Nature and numbers of variables are also necessary information

Except for variables name and context, nature and numbers of them are also necessary information for choosing the right statistical test. 'Other details' column of variables in Table 4.2 including nature of the variables, namely whether it is an interval variable, ordinal or categorical variable, and whether the variables are dependent/matched or independent groups. However, only some amount of the questioners provide the character information of the dependent variables. Even among the three of them, respondents requested for this important lacking information, the askers still didn't provide it at the end.

Besides variables information, questioners also provide other units of experiment design. As can be seen from Table 4.2, nearly half of the questions were provided with sample size. Some of the questions also provide information related to the experimental procedure and whether it's a within/between-group design.

Goal of analysis & Hypothesis

One dataset can realize different research purposes with different statistical tests

Which statistical test to use for the analysis depends upon the objective of the study, one dataset can realize different research purposes with different statistical tests. Most of the data workers provide their analysis goal in the questions. Here is one example question with specific analysis purpose:

"[...]I want to conduct a test to see if there are any significant differences between the quartile groups and the independent variables." - R2

However, just one-third of the questions have hypothesis statements. A hypothesis is a tentative statement about the relationship between two or more variables. It is a specific, testable prediction about what you expect to happen in a study. For example, a study designed to look at the rela-

tionship between emotional expression and outcome score has the following hypothesis statements:

“[...]I have two hypotheses: 1. Direct expression will lead to a higher outcome than indirect or no expression. 2. Indirect expression will lead to a higher outcome than no expression.” - C21

It's necessary to note that a question with hypothesis statement is also judged with analysis goal. On the contrary, a question that contains the analysis purpose does not necessarily also includes hypothesis statements. Because a good hypothesis will be written as a statement or question that specifies the dependent and independent variables and also the prediction of the effect.

Hypothesis statements have more strict requirements than analysis purpose

Software

I think that the second interpretation is more accurate. However, I don't know how I can check whether the change of score from time1 to time2 between different groups is statistically significant.

Thanks.

ANCOVA Syntax **SPSS** Statistical Significance Inferential Statistics ANOVA
Repeated Measures

Figure 4.1: Software tag on ResearchGate (R37)

Week	Thriller	Romance	Fantasy	General Fiction
1	19	26	10	0
2	14	24	15	2
3	13	22	15	5

Can I use these score points for hypothesis testing? Or how else can I use this data to make a comparison between different genres?

time-series hypothesis-testing statistical-significance **spss** analysis

Figure 4.2: Software tag on CrossValidated (C14)

Although we don't focus on the questions related to the implementation of statistical tests, there is still twenty percent of the questions including software information. Data scientists provide software information in two ways.

One is in the question description. In this scenario, questioners explicitly mentioned what software they already

Software information is provided in question description

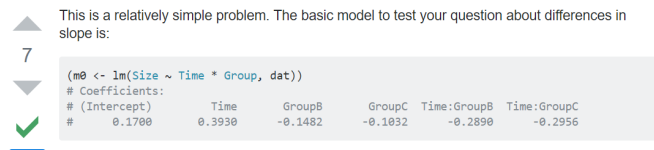
used for the study and expect respondent give them suggestion or material based on this software, here is one example:

“I am doing a mixed ANOVA in SPSS where I am currently comparing the duration of certain script handwriting in 2 conditions...” - C3

Software information is provided with tags

The other way is using the tag function provided by the Q&A websites. In these questions, data analysts don't provide software information in the question description. Instead, they implicitly choose a software tag to express their expectation of the respondents to give suggestions based on this software. Figure 4.1 and 4.2 show two example screenshots with software tags.

For these seventeen questions with software information, only five respondents give answers based on the same software, they could either be implementation information with code script:



The screenshot shows a forum post with a question and an answer. The question asks for a basic model to test differences in slope. The answer provides an R code snippet and a table of coefficients.

```
(m0 <- lm(Size ~ Time * Group, dat))
# Coefficients:
# (Intercept)      Time      GroupB      GroupC  Time:GroupB  Time:GroupC
#      0.1700      0.3930     -0.1482     -0.1032     -0.2890     -0.2956
```

Figure 4.3: Answer with software code scripts on CrossValidated (C12)

Software is implicit requirement for implementation information

Some respondents don't provide detail implementation codes, instead, they provide some useful materials or examples with tutorials on how to do suggested statistical test with the expected software. Here is one example:

“[...] You would use a dependent test because the same person rated both products instead of having two separate groups for each variable. Here is an SPSS tutorial for it: <https://statistics.laerd.com/spss-tutorials/dependent-t-test-using-spss-statistics.php>” - C31

However, not all the respondents give suggestions based on expected software. In one example, questioners already started the initial analysis with SPSS, but the respondent

mentioned the complexity of implementing the test with SPSS and provide examples link of Stata:

"[...]AFAIK, SPSS has no easy, straightforward way to get the interaction contrasts Stephen described...I am a long-time SPSS user who has started using Stata a lot more in the last few years. For interaction contrasts of this type, Stata makes life much easier IMO. You can see some examples on this UCLA web-page: <https://stats.idre.ucla.edu/stata/faq/how-can-i-explore-interactions-using-the-contrast-command-stata-12/>" - R37

In the other question (C15), respondent ignored the tag with software information 'SPSS' and gave suggested test with R code scripts. The reasons could be the test is easy to implement with R, or the respondent is more familiar with R than spss. No matter what reason it might be, the long answer with detail explanation and code scripts was not accepted with a checkmark by the questioner, which make us wonder if it's owing to unsatisfied software information.

Respondents may ignore the software information

4.2.3 Problems in Questions Formulation

The third theme centers around the statistical questioners' problems in expressing information needs occurred in two Q&A websites of our surveys. Our research process including detail content analysis of both questions and their answers, we want to find out not only how data analysts ask statistical questions and provide information, but also how respondent answer these questions with provided information.

During this analysis process, we found information gaps between question and corresponding answer. Respondents always can not give proper suggestions because they don't have enough information. Sometimes even the necessary information is available, they still confused about it because of the unclear description. The information gap was classified into two main categories: Missing information and Unclear information based on respondent comments and discussions with questioners.

Missing information and unclear information are two main problems of question description.

Missing Information (29/76)

One-third of the questions in our analysis lacks necessary information for choosing proper statistical tests.

As discussed in the previous section, the decision for a statistical test is based on the scientific question to be answered, the data structure and the study design. But what information is needed depends on different circumstances, sometimes all the information mentioned above is necessary, sometimes just some of it is enough.

One judging
criteria:whether there
is more information
required by
respondents

Based on this complex situation, we did content analysis for comments and answers below each question, the missing information can be distinguished according to two standards. The first one is finding whether there is more information required by respondents. In one question, the data worker had data consisting of body mass measures in four groups, each treated with different chemical and one being a control group and asked for the proper statistical test used for the data. However, there is no information related analysis purpose of the study. In the comments, one respondent required:

“What is the research question? Is there any statistical hypothesis related to it?” - C8

The other judging
criteria:whether
respondents assume
some information or
they gave different
options because of
lacking information

Another standard for judging missing information is whether respondents assume some information or they gave different options because of lacking information. For example, in one question, the questioner wanted to conduct a questionnaire that measures five domains in males and females, detail background information was provided except for data type which was requested by the respondent later:

“First, specify whether the data produced by questionnaire is continuous or categorical. Chi-square is a better option for categorical data. If continuous, then you need to specify whether your data is parametric or non-parametric. . .” - R20

Based on this response, data type and assumption are treated as missing information.

Table 4.7: Summary of missing information

Type of missing information	Frequency
Data type (raw/summary data,variables nature)	14
Analysis goal/Hypothesis	10
Assumption	9
Variables	4
Experiment design	4
Sample size	1

As can be seen from the Table 4.6, information of data type, assumptions and analysis purpose are the three most easily overlooked information by questioners, which to some extent also reflects that these three types of information are most important and non-negligible when selecting the appropriate statistical tests. There are also some respondents asking for variables and experiments information for making the decision. Detail reasons and analysis of why data analysts always overlook this information will be discussed in Chapter 5.

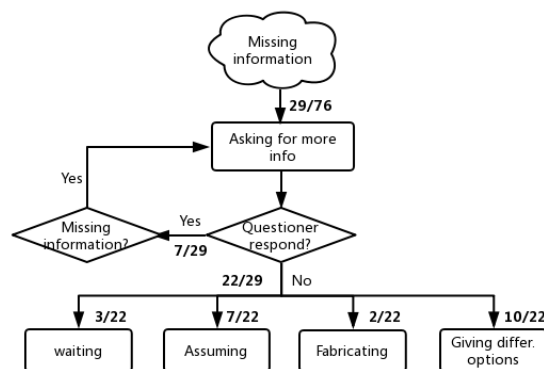


Figure 4.4: The process of respondents handling missing information

Figure 4.4 illustrates how respondents handle the questions with missing information. The ideal situation is the questioners quickly response to the request and supplements the missing information. Seven out of twenty-nine questions with incomplete information have supplementary information by the questioners. However, even the informa-

Respondent handle missing information in different ways

tion is added, two of them still were not described clearly and one of them don't get a response from any respondents later maybe because of the long-time period, below is one example question with the above situation:

"Thanks for providing more info, Jane. I suspect the distribution of hospital costs is positively skewed, is it? If so, it is likely that for descriptive purposes, one might decide to use..."-R6

Respondent
fabricated missing
information

Most of the missing information is not added after some time period, then respondents handled this problem in four different ways. In order to explain the analysis process more clearly, some (2/27) respondents fabricated the dataset with code scripts to answer the question because of no provided dataset:

"Illustrations with fake normally-distributed, paired data: The first 6 of n=128 values of T1, T2, and D=T2T1 are shown below..." - C15

Respondent assume
missing information

Some of them (7/27) assumed the missing information. In one sample question, respondent suspected the hypothesis and give proper test based on his suspicion:

"I presume that your hypothesis concerns whether the chemical leads to lower or higher mass at follow up. To test this hypothesis, you can use an ANCOVA." - C8

Assuming
information is
dangerous and could
make the answer
incredible

However, one dataset can realize different research purposes with different statistical tests. In another case shown in Figure 4.5, the respondent even did not mention he presumed the assumption information, until another respondent found the problem and give the comments. Imagine what if Martijin did not find the problem and the questioner directly used the t-test for the study, this would cause a wrong analysis result and influence for the whole research. So that presuming is always not an ideal way of handling missing information, which may lead an unsatisfied test and make the analysis in an unexpected direction.

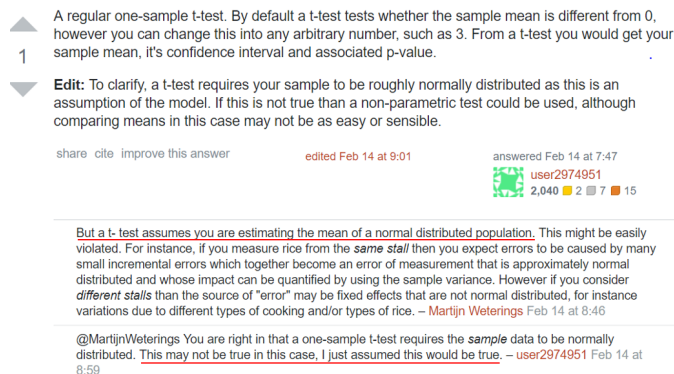


Figure 4.5: Respondent assume the missing information(C6)

Some of them (10/27) gave several options based on different situations. For example, in one question, questioner just mentioned the outcome will be scored, but what kind of score is overlooked, then the respondent gave two tests for each type of data:

“There are two types of variable, continuous and categorical. If you want to predict a categorical variable for example gender (Male and Female) so you must use logistic regression otherwise multiple linear regression for continuous variable” - R23

Compared with the previous two reactions, this way of handling missing information is more reasonable and reliable. Questioners can later choose the most appropriate one according to their specific situation.

The rest of them (3/27) of them did not react after requesting more information, they just wait for questioners additional information and did not give the suggestions until they get enough information.

Unclear Information (13/76)

One-third of the questions in our analysis have unclear information for choosing proper statistical tests. The difference between unclear and missing information is, the previous one has the necessary information which it's not understandable for respondents, while the latter does not have the information at all. If the information is unclear because lack of explanation and detail information, then it belongs to missing information.

Respondents gave several options based on different situations

Questioners did not provide required information

Edit function give us chance to retrieve improper information

The standard for judging unclear information similar to how to distinguish missing information. Content analysis for answers has been conducted and found out whether the information is confusing and queried by respondents. Furthermore, CrossValidated platform has a function which allows other experts to edit the original questions. This gives us the opportunity to analyze improper information provide by questioners. Detail explanation will be described based on two main causation of unclear information emerged in our analysis: improper content expression and unclear knowledge cognition.

Table 4.8: Summary of unclear information

Category	Sub-category	Frequency
Improper content expression	Chaotic structures	4
	Inaccurate phrase	3
	Ambiguous sentence	3
Unclear knowledge cognition	Improper analysis procedure	2
	Improper tag	8
	Wrong charts	1

1.Improper content expression

There are three types of improper content expression emerged in our analysis: Chaotic Structure, Inaccurate Phrase, and Ambiguous Sentences.

Chaotic structure

How questioner provide information influence the quality of the answers

As discussed in the first part of this chapter, a good question asking the selection of statistical tests always contains multi-dimensional information. The type of statistical test to be applied depends mainly upon the hypothesis to be tested, type of data used in testing, type of sampling design used in collecting data and the parent population from which the sample is drawn. After discussing 'what' information should be included, in this part, we will discuss 'how' the information is structured by questioners.

Many questioners described their questions with clear information titles which indicate each item clearly, from the descriptions, respondents can clearly judge every informa-

tion clearly, here is one example with which listed each information with indicators:

"[...] I have a DEPENDENT VARIABLE which is a diet quality score that is based on quartile groups so ordinal (quartile 1, quartile 2, quartile 3, quartile 4). My INDEPENDENT VARIABLE is BMI, Weight, Waist circumference so continuous. I want to conduct a test to see if there are any significant differences between the quartile groups and the independent variables. Which statistical test should I conduct?" - R2

One questions provide clear information titles

However, not all the information was provided in a clear manner, some questioners provided information in a hybrid way. This hybrid information always mixes available matrix together and make some information implicit, which is inconvenient for the respondents. In one question, the asker used word 'column' to describe both variables and levels and didn't explicitly indicate what are independent and dependent variables:

Many information is provided in a hybrid way

"[...] I've got 3 columns one for pre manipulation of language which will be my control. I only have positive and negative columns. In addition, I will also have a column for age(3 levels) and gender (2 levels) and want analysis how these demographics affect the effect that positive and negative language has on investor preferences." - R19

As the consequence, two respondents restructured the question again to verify his understanding and then gave the suggestion, compared with the original version, the respondent listed all variables separately and give a clear analysis question, here is how the respondent reformed the question in the response:

Respondents restructured the question again to verify his understanding

"[...] 3 columns: 1- language classified into +ve -ve. 2- age classified into 3 levels. 3- gender classified into male female. And I understood that: investor preferences is the outcome Question please: Is the var. investor preferences is qualitative?" - R19

Twenty-four Questioners provide their datasets with charts. Compared with file, people prefer tables and graphs since they present the dataset in a more 'visualized' way.

People prefer provide dataset with charts

However, this way in the Q&A websites also has some limitations which caused unclear information.

Tables without values

Three tables (R1, R14, R16) are provided in the questions were without values. For these questions, table is just a function to present variables and loses the original advantage of visibly showing the data. Extra description of about the datasets such as whether it is continuous, ordinal or categorical, is still necessary.

Table without column and row names

One table was provided without column and row names (C45). Without this information, pure data is meaningless. In the question, just frequency table was shown with no additional description which was confused to respondents and it's obvious that they asked the meaning of the values later in the comment:

“You need to explain the data structure more, otherwise it is not clear how would a chi-square test apply. What do the values represent? Are the rows meaningful?” - C45

Table with unclear column and row names

One table has unclear column names which were required further explanation by the respondent(C7). The column and row names normally reflect the variables in the study. However, because of the table's space constraints, a detail description of the variables is not allowed. In the question, questioner just uses 'case' and 'control' as column names to describe two groups of people and lack context information of the variables. The respondent later asked the difference between the two groups:

▲ Definite stats newbie, and I'm looking for a little bit of direction on which statistical test to use.

0

	Case	Control
> 1 Narcotics	1304	133
<= 1 Narcotics	6095	3404
	7399	3537

▼ Here's a quick summary:

★ I'm trying to determine if the cases are on more than one narcotic at a time with a .05 significance. Is this an easy as doing an OR $(1304*3404)/(6095*133)$?

hypothesis-testing | t-test | odds-ratio

share cite improve this question

asked Mar 19 '14 at 14:07
 wootscootinboogie
 100 8

what's the difference between the case and the control? it looks like the control is also on narcotics. – markovchain Mar 19 '14 at 14:20

@markovchain The cases are people who have sustained a particular type of injury, and the cases have not sustained that injury – wootscootinboogie Mar 19 '14 at 14:25

Figure 4.6: Problem with table structure in the forum (C7)

Two charts only showed part of the dataset (R21, C13). Sometimes the amount of real data is too large to present all of them in the question space, so that some questioners just show a small part of it to describe the question. However, this could cause a wrong choice of the statistical test due to incomplete information. For example, if the questioner lacks experience, only part of the skewness data is shown and the tail part is overlooked by respondents, they will misunderstand the data as normally distributed and give wrong suggestion.

Table has space limitation to show all the dataset

Inaccurate Phrase

Three questions were provided with inaccurate phrases (C4, C6, R15). Like every subject, statistics has its own language. The language is what helps you know what a problem is asking for, what results are needed, and how to describe and evaluate the results in a statistically correct manner. An inaccurate phrase or terminology in the description will mislead the respondents. For example, in one question, the phrase 'in vitro data' was confusing and asked by the respondent later:

Statistics has its own language

"Suppose I have an in vitro data of three biological replicates for a comparison between two independent groups, presumably a two-tailed test since I want to investigate for any difference." - R15

In another question, questioner used 'theoretical values' to describe the dataset, but the meaning of this phrase was asked by the respondent later:

"[...] we have these theoretical values (in Calories) for our rice samples" - C6

In order to reply to the question and give a proper suggestion, respondents assume the confusing phrase with his/her own understanding.

Respondents assume confusing phrase

Ambiguous Sentences

Three questions were provided with ambiguous sentences (C1, C16, R36). Ambiguity is not just happened in the statistical description. The definition of ambiguous is something that is unclear or not easily describable and

Ambiguity is not just happened in the statistical description.

sometimes is also used deliberately to add humor to a text. However, in statistical question description on Q&A websites, ambiguous sentences always confuse the reader and hinder the meaning of the text. In one example, the questioner illustrated the hypothesis as follows:

“My hypothesis is: “Regardless of the context, men and women differ significantly in emotion expression...” - R36

One ambiguous hypothesis statement with two interpretations

this hypothesis caused different interpretations among the respondents, one of them said it could have two meanings based on whether the variable ‘context’ can be ignored or not, and gave two possible answers for different interpretations separately.

In another question, the asker described the experiment as follows:

“Each subject is assigned to one of three different Times (morning, noon or afternoon) and receives one Treatment (A,B or C).” - C1

One ambiguity causes unclear experiment design

With this description, one respondent thought it was designed in a repeated-within-subject nature and give the corresponding solution. The misunderstanding was later corrected by the questioner and made it clear that each participant was only tested for one time with one treatment.

As can be seen from the previous two examples, ambiguity in statistical questions is not as funny as in normal day life and could mislead the answers in the wrong direction.

2.Unclear knowledge cognition

The other type of unclear information is caused by a lack of cognition, improper Analysis Procedures, Tags and Charts are used to present the information.

Improper Analysis Procedures

Two questions in our research used improper procedures to analyze their data set initially. Data analysis procedure involves multiple steps after getting the data, normally initial data processing is necessary.

In one question, the questioner created a point system for

the best seller's list and added up ranks column, which not only messed up the data but also confused the respondent:

Questioner wrongly processed the data

"I'd question your method of creating points. Why should ranks be additive? The difference in sales between rank 1 and rank 2 may be very different from that between rank 9 and 10. And the difference between 1 and 2 in week 1 may be very different from the difference in week 2. So, that's going to make the data very messy." - C14

In another question, the asker wrongly used the Kruskal-Wallis test for the dataset which violates the assumption and asked by the respondent later:

Questioner used the wrong test

"If each row represents a region and they are not ordered, then how are you doing a Kruskal-Wallis test? That test only applies to ordered outcomes." - C25

Improper Tags

A tag is a keyword or label that categorizes the question with other, similar questions. Using the right tags makes it easier for others to find and answer the question. CrossValidated platform has the 'edit' function, which provides the opportunity for both questioners and respondents to edit the questions and tags. Based on the editing feature, we analyzed what information has been edited later by experts.

Tag function is important in filtering and recommending topics

The result shows that eight out of thirty-six questions posted on CrossValidated contains improper tags information: four of them with wrong tags and the rest four missing proper tags, for which experts later either added or changed the tag to increase the match and the probability of being answered.

Questioners have problems of choosing proper tags

Wrong Charts

One questioner used the wrong graph to present dataset(C30). Not all the charts and graph are good at visualizing the data in an informative way, it depends on the data type and distribution. Choosing the wrong visual aid or simply defaulting to the most common type of data visualization could cause confusion with the viewer or lead

to mistaken data interpretation.

Edit: It now appears you do have a count. This is critical information and should have been made clear to start with. Your histogram is both not very informative and actively misleading, so it helped lead me astray. I'll come back and edit some more, but for now, some advice (besides "explicitly mention when your data are counts"):

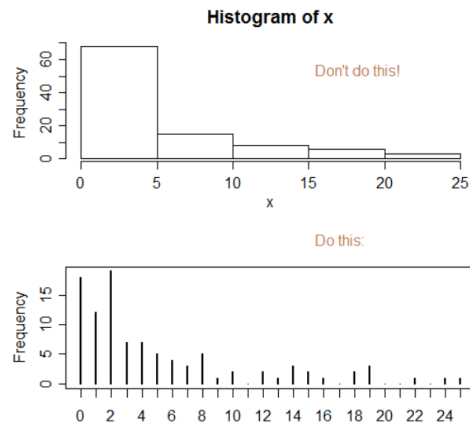


Figure 4.7: The wrongly used chart on the websites (C30)

one question with
wrong visualization
chart makes
respondent confused
about data type

In this question, the questioner used the histogram to show counts and did not explain the data is count, which misled the respondents. Detail screenshot is shown in Figure 4.7.

As can be seen from the previous examples, questioners inappropriately used analysis procedures, tags, and charts because they lack some specific statistical knowledge. The cognition misunderstanding causes unclear information provided in the question space.

Respondents handle
unclear information
in different ways

Figure 4.8 shows how respondents handle questions with unclear information. Most respondents perceived the unclear information in the questions' description, except for one respondent (C1) thought the experiment as repeated-within-group design because of an ambiguous sentence in the question description. Fortunately, the misunderstanding information was then corrected by the asker and did not make a further bad effect on the study.

Two out of thirteen questions with unclear information have supplementary information by the questioners later. Most of the unclear information was not explained after some time period, then respondents handled this problem in three different ways.

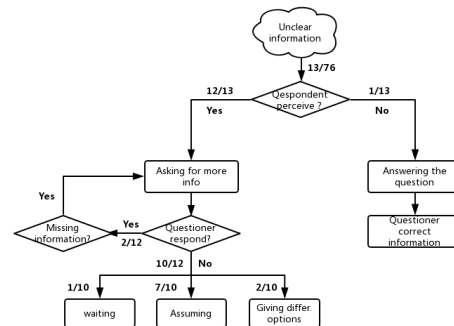


Figure 4.8: The process of respondents handling unclear information

Most of them (7/10) assumed the unclear information. In one sample question, questioner described the experiment design with an ambiguous sentence which can be understood as both between and within-group design, then the respondent suspected the information:

“[...] I’m assuming that both algorithms worked on the same set of documents and that the data is paired.” - C16

However, assuming is always not a proper way to handle unclear description. It is ideal that respondents state what information is unclear and being assumed, if it is not explicitly stated, the questioner will accept the unreliable answer and make a wrong result.

One respondent (1/10) did not react after asking for an explanation of the unclear phrase: ‘theoretical values’ in the question, he/she just wait for questioners additional information and did not give the suggestions until detail explanation is provided.

Some of them (2/10) gave several options based on different situations. For example, in one question, questioner presented the hypothesis statement with an ambiguous sentence which could have two different interpretations, then the respondent listed both situations and gave answers separately:

“Sandra, I have two interpretations of your hypoth-

Respondents assume unclear information

Assuming information is dangerous

Respondents give different options based on different situations

esis: 1. I can forget the variable "context". In this case, I ignore this variable and a two independent samples T-test is adequate to compare the two genres. 2. I can not forget the variable "context". In this case, I first eliminate the effect of "context", that is, I apply an analysis of covariance, taking "context" as a covariate."-R36

Compared with the previous two reactions, this way of handling unclear information is more reasonable and reliable. Questioners can later choose the most appropriate one according to their specific situation.

4.2.4 When Do Data Scientists Choose Statistical Tests?

A successful study should include many continuous steps

A research study involved many steps, which could be summarized as a six-steps process shown in Figure 11. Careful statistical planning, including the selection of study endpoints, the determination of the required sample size, and the selection of statistical tests to be used in the data analysis, is important to ensure a successful research project [35].



Figure 4.9: The process of a research study

The timing of choosing statistical information throughout the whole research process

As stated in the beginning, our analysis focuses on the questions asking about choosing statistical tests. However, after close analysis of these questions, we found that the timing of asking for statistical information throughout the whole research process. Generally, question timing can be summarized in three phases: Before Data Collection, After Data Collection, and After Initial Data Processing.

Question timing is judged by provided information

One classification criteria is based on the provided information. If raw data or some initial processing result is provided, we can easily judge the question timing is after data collecting or data processing process. The other standard is based on the question description. For example, in one question(C11), the questioner said: "I'm designing

an experiment to evaluate reagents used to develop fingerprint...”, which implicitly reveals the experiment is still not conducted.

With the two standards mentioned above, most questions can be classified accurately, except for nine questions. In these questions, either data or clear description provided to distinguish whether the experiment is already conducted. Here is one example:

“I am trying to study the association between blood groups (categorical variable) and cholesterol levels (continuous variable) adjusted for confounders? I was wondering which statistical test would be more appropriate? Thank you beforehand for any assistance.” - C5

In this question, variables and the analysis question are clearly illustrated, but only with this little information, we cannot judge if the question timing is before or after the data collection process.

Before Data Collection (4/76)

Only four questions were asked before the data collection process. These questions are not only related to statistical analysis but also require information in previous stages, such as study design and sample size. In many ways, the design of a study is more important than the analysis. A badly designed study can never be retrieved, whereas a poorly analyzed one can usually be reanalyzed. Consideration of design is also important because the design of a study will govern how the data are to be analyzed.

An important advantage of consulting a statistician before a study begins is that she can verify that the planned procedures and size of the study will be adequate to address its goals. In one example, the questioner clearly stated the research question but still did not decide the study design and sample procedure, which reveals that he/she is still on Step1 shown in Figure 4.9:

“My research question is comparing nursing students knowledge for those who have had a medical terminology course compared with those who have not. I would like to utilize a question-

Nine questions cannot confirm the question timing

Also require information in previous study stages

The expert can involved in previous stage to avoid design mistake

naire/survey to obtain this information. Where I am confused is what type of testing would this questionnaire/survey be (t-test, ANOVA, etc..)" - R9

One question asks about questionnaire design

For this question, respondents gave suggestions on how to construct the questionnaire and also which statistical test to use. In another example, the questioner was working for a company which helps a group of students to prepare for a standardized test and wanted to know how to conduct a study with a general goal. It is obvious that the questioner had no clue with study design and also sample procedure:

"Here are my questions:1. What sampling procedure should the company use? 2. What statistical tests should the company apply to the data to see whether there is a statistically significant difference in difficulty between tests?..." - C27

One question asks about study design and sample procedure
Answers provided before data collection is more credible and reasonable

For this question, the respondents give suggestion from study design to statistical analysis to help the questioner conduct the research step by step.

Questions asked before data collection are always accompanied by other requests in previous stages such as how to form a good research question, how to design the study, or how to estimate the sample size. For these questions, expert respondents will give suggestions from research's early stage to the final stage which is more credible and reasonable.

Asking questions without dataset could cause unclarity

However, asking questions without dataset could make cognition gaps between questioners and respondents, because of fuzzy expression and lack of information. As discussed before, for handling unclear and missing information, respondents may assume the information with their own understanding of the situation, which could cause a wrong answer. Here is one example comment below a question without data:

"Can you say how the data are intended to provide information about a connection between working memory and problem-solving ability? If those are connected with Tasks 1 and 2, can you say how? In plain English, can you state your objective in terms of the two tasks? // Can you show the data for at least a few of the subjects?" - C15

The unclarity happened because questioner didn't describe the experiment properly, but if the data was shown here, unclear information can be inferred from the data and the problem will be solved effectively.

After Data Collection (36/76)

Nearly half of the questions were asked after getting the data. In this stage, the type and nature of the data are clearly known and can be provided in question space to help the respondents have a better understanding of the situation.

Compared with asking before getting the raw data, detail information about raw data can be provided. Data information can make up for fuzzy expression. Sometimes a well-structured dataset is worth a thousand words. In one example, the questioner showed two rows of real values getting from the experiment and asked for verification about choosing the statistical test. With the provided data, the respondent clearly found the data is very close to 1 and also highly skewed, then gave the proper suggestion based on this character:

Detail raw data information can be provided

"Not only do you have heteroskedasticity, with proportions (as in your example) very close to 1 (or to 0) they'll also be highly skewed (e.g. the individual values can't exceed 1, but for a value close to 1, they could easily be more than that distance the other side of the mean). High skewness can have a bad impact on t-tests..." - C10

However, choosing statistical tests after experiments also causes troubles. For the novice, if there is a problem with the initial experimental design and sampling process, then the data collected will be problematic. In one example, the questioner already gathered data from fifty-seven questionnaires then asked which statistical test to use for analysis on Research Gate. However, the respondents pointed out the irrationality of the questionnaire design and suggested to generate more questions:

choosing statistical tests after experiments also causes troubles

"If your dependent variable consists of only a single item with 5 points Likert scoring, then this will limit you to using non-parametric statistics. I suggest that

you try to generate other similar questions that you can combine into a scale, because that kind of continuous measure would allow you to use a more flexible set of statistics.” - R7

The questioner replied to accept the suggestion but not sure whether her industry mentor would allow her to increase questions as questionnaire almost finalized.

Having the risk of the previous steps being wrong

As can be seen from the above example, choosing statistical tests after collecting all the data could have the risk of the previous steps being wrong. If all the experiment has been finished and found that the original design is not suitable, then it will consume a lot of resources to start from the beginning.

After Initial Data Processing (27/76)

More than one-third of the questions were asked after initial data processing. Presenting these processing results can help questioners better understand the nature of the data.

Processing is a series of actions or steps performed on data or descriptive coefficients

The processing could be a series of actions or steps performed on data to verify, organize, transform, integrate, and extract data in an appropriate output form for subsequent use. For example, one questioner (R21) already did the Shapiro-Wilk normality test and F test to verify the distribution and ratio of variances. Another questioner (R5) did the Discrete Cosine Transform(DCT) to the original data and want to verify the proper statistical tests. The processing could also be descriptive statistics which are brief descriptive coefficients that summarize a given data set. Such as in one example question(R8), mean and median were provided.

These initial data processing procedures help respondents know the detail information of the data and give more credible answers.

Chapter 5

Discussion

The main objectives of this research study were to find out how practitioners seek statistical analysis procedures, how they search useful information to choose the proper statistical test, what resource they use and what are the general problems they have met. In this chapter, an interpretation of the findings and possible reasons for them will be discussed. We compare the different aspects of our findings and results to prior work and relate the findings to other research carried out.

5.1 Possible Causes of Uncertainty

The first theme of the interview results reveals that practitioners feel uncertain when choosing the statistical test and have used wrong methods for their analysis. This theme directly correlates to the previous research results which show issues in significance test selection and confirms this is a general problem of data scientists [12, 30]. Why choosing the right statistical test is a general problem of data workers? I discussed this with the participants and find the following three possible reasons.

The first one could be lack of the standard. Although there are lots of materials and resources try to help data workers make the right decision, such as some online pages [26] and research papers [50], there is still not a standard and commonly used resource which could help data workers solve this problem:

There is no standard
in statistical test
selection

“There is no standard about it. It’s hard for researchers to know which is the right test to use and how to use it”-P04.

They search for available information from different resources, compare and combine all of them then make the final decision. One participant described this complex procedure in the interview:

“I googled statistical test, first of all, I checked what tests are there, it’s a kind of overview, for every test I checked again google or some books what the test tells about. For example, I usually go for t-test examples and check out what they did, does it make sense for my analysis”-P09.

People lack targeted statistical education

Secondly, the lack of adequate statistical education is another main reason. It is well recognized within psychology and social sciences that there is a poor understanding of significance testing amongst students, teachers, and researchers [37]. Most of the participants are not experts in statistics, they come from different research fields, such as HCI, psychology, and the economy. They only use statistical methods to analyze the data in the experiment, so the frequency of using statistics is relevant slow. Most of them have only had basic statistical analysis courses during Bachelor, and have not accepted systematic training, so they lack practical experience in statistical analysis. Some participants indicated the gap between practice and course knowledge in the interview:

“We have SA course, it’s really basic and similar to what I learned by myself before. It’s not practical.”-P05

Choosing statistical test is a complex procedure

Thirdly, situations change every time according to experimental design and characters of the data. There are many factors influence which statistical to use, the most important of which are data characteristics and research purposes, which are determined by experimental design. Even a little change could make the final decision different, the logic behind choosing the right test is really complex.

Uncertainty and wrongly use of statistical test can cause serious problems

Statistical inference involves tests of hypothesis, where statistical tests play a crucial role. Statistical methods are required to ensure that data are interpreted correctly and

that apparent relationship are meaningful and not simply chance occurrences. Uncertainty and wrongly use of statistical test can cause serious problems. For example In the field of medicine and nursing, if the researcher makes mistakes in calculation while performing the statistical tests, then the researcher might end up committing a type II error. In other words, the statistical tests will conclude that a false drug sample is a correct drug sample. Further, the researcher might end up tagging a false drug sample as a correct drug sample. Thus, the researcher should be cautious while performing statistical tests. In the field of medicine and nursing, errors in statistical tests can result in huge problems in people's lives, as it affects their drugs and dosages, etc.

5.2 Reasons for the Subjective Attitude and Existing Problems

The second theme shows some interesting phenomenons of how data workers choosing statistical methods. We do not focus on the objective but subjective factors, such as psychological activities hide behind the surface and have a huge influence on making the decision of statistical tests.

As can be seen from these findings, data workers prefer using the most common methods while holding negative attitudes toward the new method. Pet et al [54] reviewed papers published in basic medical science research journals (included Anatomy, Physiology, and Pharmacology) and found that authors heavily relied upon the application of well-established statistical methods only and it seems that they are avoiding using new statistical methods, which is consistent with our finding. Nour-Eldein H [51] also indicated that basic analyses were used slightly more in articles than advanced analyses. The sophistication of statistical methods are going to be increased over time and avoiding the use of advanced techniques may miss many possible important inferences from the same data.

We will discuss possible reasons in the following part. Firstly, publication concern will affect their choice of statistical tests. Publishing papers has always been one of the biggest sources of pressure in researchers' academic

Studies show that basic analyses were used slightly more in articles

Researchers have publish concern

life. There is a variety of research have shown that publish concern is like a double-edged sword, which can have an impact on research in all aspects. In our interview process, many researchers showed different levels of anxiety about publication. This anxiety influences their decision of choosing the statistical method for their studies. Many researchers mentioned that when they choose the statistical method, they will consider whether the reviewer of the paper understand and accept the power of this statistical method. Two HCI researchers talked about this in the interview:

“There are also other thought such as if I use this method in my paper, will the reviewer of the paper know this approach, if they won’t will they reject my paper? so normally we will just stick to the methodology already out there, it’s a common thing. ”-P04

Using common method saves time for both authors and reviewers

It seems that using common statistical methods in their field is a win-win approach for authors, editors, and reviewers. All of them do not have to spend extra time learning and validating the reliability of the new method. So that papers with commonly used methods need shorter time for reviewing and are published sooner which motivate researchers to use common methods more often. This closed-loop explained why researcher who with publish pressure prefer choosing the most classic and traditional statistical tests in their study.

Industry people emphasis more on the explanatory nature of the model

For people working in the industry, things could be different. One main concern could be the model’s interpretability. Hand, D [38] mentioned this concern in the banking industry. The term ‘front end’ in banking industry refers to models in which the result is communicated to the customer. In situations where the application will be told the outcome such as scorecard, there are often legal obligations to tell the customer on what basis a rejection has been made. This means that such models must be interpretable so that sophisticated neural network or support vector machine models will not be appropriate. Except for communicating to customers, colleagues and bosses understand the meaning of the model is also important in the work. One participant working in the economics industry mentioned this in the interview:

“when we finished our model we have to report layer by layer, first is the team manager, then is department manager, why you use this indicator in your model to predict the risk, it should be accountable”-P07.

He said this is the reason why the regression model is the most frequently used statistical model in his working field (risk analysis) which consistent with Hand, D’s finding that most popular tool for scorecard construction is logistic regression.

Another main influence factor in industry field could be time limitation. Rapid processing and decision-making are often required, sometimes in real-time, such as deciding whether to give a loan to a customer. But some complicated models such as customer segmentation which uses cluster analysis may be carried out for weeks. One participant who has talked with a Sweden bank company provided that:

“I asked if they use deep learning model for analysis, they said no, we don’t use that complex method, we just use simple logistic regression even though former one could get a better result. Because they have a really huge amount of data, every week they running their model, if the model is too complex is time-consuming”-P08

Time issue also limited industry people choosing statistical tests, which result in models having rapid running time such as regression win the game in the industry field.

However, sticking to the most commonly used statistical method or a familiar method could raise problems. Kaptein M and Robertson J [36] mentioned that researchers from HCI, psychology, medical and economics typically use a significance testing approach to statistical analysis when testing hypotheses during usability evaluations. But the controversy over this traditional method has lasted for over forty years since Cohen first noted it in 1994 [18]. Critics of the traditional statistical inference method of significance testing argue that “it is time for researchers to consider foundational issues in inference” [22]. Similarly, Wagenmakers, et al. conclude that “experimental psychologists need to change the way they conduct their experiments and analyze their data” [62]. There are lots of other

Running of complicated models could take long time

Researches show different arguing against the use of some traditional methods

research shows different arguing against the use of routine interpretation of results using p-values. However, researchers, reviewers, and editors still think the result is statistically significant or “publishable” when the p-value is less than 0.05. In school, students still learn this as a golden rule for getting a significant result. Except for this p-hacking issues, avoiding using modern method and sticking to the traditional statistical method may miss many possible important inferences from the same data.

5.3 Types of Questions About Statistical Test Selection

Data scientists who ask questions on the topic of choosing statistical tests in both ResearchGate and CrossValidated are seeking to fulfill a tangled web of needs, sometimes all bundled within the same question. Based on the information requirements, all the questions can be classified into three main categories: seeking factual information, seeking validation, and seeking resource.

There are three possible reasons of people asking statistical information on Q&A sites

The primary motivations for asking statistical tests questions are to fulfill cognitive needs, more specifically, to fill the knowledge gap. Data scientists tend to expect that they will not only acquire information related to which statistical test to choose, but also how to plan the experiment procedure and how to interpret the result. It is not surprising that the motivation of acquiring knowledge information plays an important role choosing to use Q&A website since these services constitute a knowledge exchange community [1] that facilitates a user-driven environment for information seeking and sharing [16].

Answer from an expert is more credible than the random results of a search engine

It was curious that many of the fact-based questions which cover different aspects and stage of the study process, could have been answered by searching the web with a search engine, or asking real-life domain experts, which leads us to wonder why data scientists bothered to post their questions in Q&A websites. This speaks to several issues, one of which is information credibility. Data analysts questioners may think that an answer from an expert in the statistical field is more credible than the results of a search engine.

Another reason could be hard finding real-life experts in a specific domain. This connects and verifies the finding from our interviews, many participants complained that they do not know where to find the domain experts to solve their problems. Almost all 'Ask the experts' Q&A websites have the function of expert recommendation. There are many studies provided quantitative measurements [49] and strategies [5] for expert searching in Q&A websites.

It is hard to find expert in real-life

The last reason could be it is difficult to refine the question to keywords and ask in search engines. The most important thing to use the search engine is to be able to refine the right keywords and not to enter the whole sentence. However, selecting proper statistical tests is always complicated and multiple information is needed to make the right decision, it is impossible to describe the situation clearly and find the right answer with only a few keywords. This leads data analyst seldom to use search engines to help them choose the right statistical methods. Compared with browsing results of search engines, users present detailed information needs, and get direct responses authored by humans.

It is difficult to refine the complex question to keywords on search engine

The second motivation of statistical tests selection questions is to acquire verification from domain experts. This again corresponds to our interview findings presented in section 4.1 Theme one, showing that statistical practitioners feel uncertain when choosing the statistical test and have used wrong methods for their analysis. Previous research also shows issues in significance test selection and confirms this is a general problem of data scientists [12, 30]. We listed and explained three possible reasons in section 5.1.

The last motivation of the analyzed questions is asking for useful materials to help them make decisions. This verifies our previous finding showed in section 4.1 theme three. Compared with searching randomly online, resource recommended by domain experts is more authoritative and efficient.

These two questioners' motivations inline with our interview findings

5.4 Reasons for Vague and Missing Information

Community-driven question-answering websites are a par-

The distribution of quality of user-generated content has high variance

ticular form of user-generated content, which allows users to participate in content creation, rather than just consumption. However, content quality has always been a concern of the user-generated content platform. A complimentary, and concurrent, the study of questions and answers quality was performed by Agichtein et al [2], and result revealed the main challenge posed by content in social media sites is the fact that the distribution of quality has high variance. Two main factors influence the post quality is vague and missing information.

Many studies and material [26, 50, 61] presented that the statistical test to be used depends upon the type of the research question being asked, the type of data being analyzed and the number of groups or data sets involved in the study. Our analysis results correspond to previous studies' findings and further subdivide the provided information to six types shown in Table 4.2. Actually, not all this information is required, some questions can be clearly answered with only two or three types of information while some could need all six types of information provided. It depends on the question type and complexity of the study.

It is also because of the complexity of the choice of statistical methods, many questions lack important information, which makes the respondents can not give the right suggestion. The causes of this overlooked information are various, we summarized all the possible reasons as follows.

Questioners do not know which information should be provided

The first reason could be lack of experience. Statistical practitioners have various levels of experience, some of them are experts in statistical analysis, while most of them are from other domains, such as HCI, psychology, and the biology, and do not have systematic statistical education. They lack experience in statistical analysis especially for beginners, so that they do not know which information should be provided. As one questioner noted in the question:

“Please let me know if you would like more detail. I’m still something of a statistics beginner, so I’m not sure how much depth is required to be able to answer this question.”-R5

Questioners do not want to provide own data because of data privacy issues

Second could be the data privacy issue. Many questioners fabricated not only raw data but also variables which have similar nature with their own study or just provide gen-

eral information of the variables, but without real name and context of the study. The reason people prefer to spend time falsifying information rather than real information is the privacy issue. With the strengthening of people's data security awareness, there are many regulations limiting data abuse and exposure. What's more, from the perspective of the researchers, they always do not want to disclose research topics prematurely to the public, so that the use of counterfeit data and study context is widely used by questioners.

Lastly, the information provided is largely determined by the question timing. As discussed in section 4.2.4, question timing can be summarized in three phases: Before Data Collection, After Data Collection, and After Initial Data Processing. If questioners consider choosing statistical tests before collecting the dataset, it is obvious that real data can not be provided at this stage. Similarly, if the question was asked before study design, detail context information about the study can not be provided in the question.

Question timing
decide provided
information type

As for vague information, main causation are the improper content expression and unclear knowledge cognition, which have been discussed as two classification standards in section 4.2.3.

It's necessary to note that questioners asked statistical questions without a formalized structure, and provided information in a hybrid way. Many people prefer tables and graphs since they present the dataset in a more 'visualized' way. However, due to space constraints. it is hard to show the complete data and variable information with just a chart, which could cause the information unclear.

The wrong use of tail tags is another cause of unclear information. The reasons for wrongly selected tags could be insufficient system design and users lack of field experience. Tags can facilitate indexing, searching, and knowledge mining of the content in Q&A sites. Under the guidance of tags, answerers can quickly locate questions within their expertise and provide answers with higher quality. This high-quality content raises the overall quality of the Q&A portals and can attract more users.

Questioners select
wrong tags

Although tag recommendation has been studied for a long time [33, 56, 24], little attention has been paid to tag rec-

Lack of research
leads to defects in
tagging system
design in statistical
domain

ommendation for questions in Q&A websites, especially for questions in the statistical area. Lack of research leads to defects in tagging system design. For example, tags in CrossValidated are manually selected by the questioners through a huge tag list, which causes missing or wrong selections for questioners, especially for beginners in the statistical field. The other website ResearchGate can automatically create tags from the question description, however, it does not allow other experts to edit the tag later.

5.5 Proper Question Timing

The analysis result showed that the timing asking for statistical tests selection throughout the whole research process. Generally, the question timing can be summarized in three phases: Before Data Collection, After Data Collection, and After Initial Data Processing.

Many experts indicate the appropriate statistical analysis should be decided before starting the study

Many statistical experts think it is important to have an experimental design planned out before the start of collecting data, and to have some an idea of planning on analyzing the data [44, 52]. Barun et al [50] stated it is important that the appropriate statistical analysis is decided before starting the study, at the stage of planning itself, and the sample size chosen is optimum. These cannot be decided arbitrarily after the study is over and data have already been collected.

However, only four questions were asked before the data collection process, the overwhelming majority of questions were asked after getting the data, which violated the recommendations of many previous studies. Why most of the questions were asked after getting the data, possible reasons could be summarized as follows.

Beginners do not know which is the proper selection timing

For some beginners, they do not have previous knowledge about when to choose the statistical test for their study. They just finished their study and found themselves without any ideas or countermeasures on how to deal with this data.

Intermediate data practitioners are more confident in selecting tests

For intermediate level statistical practitioners, they are more confident with statistical analysis and believe they can handle the statistical part after getting the data. How-

ever, different types of data are analyzed in different ways, they may find the data collected involves methods and/or responses that may be different from those to which they are accustomed.

Last reason could be a planned study takes an unexpected twist. Rather than improvise and hope for the best, people consult the experts on websites who can help to weigh the merits and drawbacks of different possible actions.

A planned study could take an unexpected twist

5.6 Potential Solutions

Based on the results of our study, we propose the following suggestions for the continued evolution of statistical Q&A websites for supporting data scientists' statistical tests selection.

The system should figure out question timing at the very beginning. We have found that data analysts choose statistical tests mainly at three different stages, before data collection, after data collection, or after initial data processing. The timing decides what and how detailed the information can be provided. It is necessary that before the user input the information, a question box with three options show up and ask the user to choose the current research stage.

The system should figure out question timing at the very beginning

According to different timings obtained from the previous question, the system should tell data scientists what information should be provided at this stage. Our analysis result shows that data scientists mainly provide six types of information to describe their problems in choosing statistical tests, but not all of them are necessarily provided. The system should present all these six types of information to data scientists first. For questions asked before data collection, at least experiment design (variables type, study procedure) and analysis purpose are mandatory, the rest information is optional to provide for questioners; For questions asked after data collection, the system should prompt users to provide different kinds of datasets.

The system should tell data scientists what information should be provided

At the same time, the system should help users to provide the information in a more structured manner. Our result shows that the provided information is unstructured and data scientists provide all the information in a hybrid way.

The system should help users to provide the information in a more structured manner.

Combined with the previous implication, the system could give six blocks presenting six types of information and give questioners the flexibility to choose and fill each block. For the dataset block, the system should also provide different kinds of charts to help data scientists provide their data in a proper way.

The system should have the fabrication function to help users fabricate datasets and variables.

The system should have the fabrication function to help users fabricate datasets and variables. One important finding of our research is that people use fake data or variables to describe the situation because of privacy issues. Based on this finding, the system should ask the user to provide the distribution and features of the data and create a dataset similar to the real one. For variables, type and number should be provided first, then the system should fabricate the similar research context and variables automatically.

The system should alarm the questioner with every comment and encourage questioners to complement the lacking information.

The system should alarm the questioner with every comment and encourage questioners to complement the lacking information. Our analysis reveals that among the questions which respondents ask for more information, only a small part have questioners response. This causes respondents to presume or give different options based on their own understanding of the scenario. In order to avoid the unreliable assumption, the system should alarm function to remind questioners of new comments and provide incentives to encourage users to complement the lack of information.

The system should automatically generate and allow experts to modify the question tag.

The system should automatically generate and allow experts to modify the question tag. Under the guidance of tags, answerers can quickly locate questions within their expertise and provide answers with higher quality. However, research result shows that many questions asking for statistical tests selection do not have proper tags. In order to solve this problems, they system should automatically generate the tags based on keywords and topic extraction and also allow experts to modify the tags later to make questions easier to understand and thus accelerate the process of the questions being solved.

Chapter 6

Conclusion

The purpose of this study was to understand how practitioners choose statistical procedures. In particular, how they search useful information to help them make the decisions.

In this study, we interviewed couple of data scientists and the result centers around three main themes. The uncertainty of the selection is a common concern among data scientists and they have used or recommended wrong answers to others. Second theme reveals subjective criteria for selecting the proper tests, they prefer to choose the most commonly used method in their research field and hold negative attitudes towards learning new methods. The last theme shows that data scientists use available resources to help them make decisions, such as books, papers and communication with others. Since all of them mentioned that they have used the information on Q&A sites ,we decided to analyze the Q&A platform to understand how they ask questions ,what information they provided and what problem they are facing.

We searched questions asking for selection of the statistical tests on StackOverflow, CrossValidated and ResearchGate and had totally 76 questions coming from different domains.

The analysis of these questions and answers shows that data scientists use Q&A forums for one of three purposes: seek factual information, seek validation, and seek resource, which verify and complement our interview find-

ings.

They provide mainly six types of information to present their problems in choosing statistical tests, it is necessary to note that some of data scientists prefer using fake dataset and variables information to describe their problem which may due to data privacy issues. Result also indicated that missing and unclear information are two main problems in question description. The reason of missing information could be summarized as lack of experience, data privacy issues and question timing. As for unclear information, it because of improper content expression, such as chaotic structure, inaccurate phrase and ambiguous sentence. Unclear knowledge cognition is another main reason, which consists of improper analysis procedure ,improper tags and wrong charts. For handling the missing and unclear information, respondents either give different options, assume or fabricated information to answer the question, which could be really dangerous and make the information incredible.

Another interesting finding is when data scientist choose statistical tests, question timing decides what information can be provided at this stage. Analysis result shows that the question timing can be summarized in three phases: Before Data Collection, After Data Collection, and After Initial Data Processing. For questions asked before collecting data, they always also ask steps before analysis, such as the plan of the study and experts and involved in earlier stage, so that the answer is more reasonable and credible. For the questions asked after data collections, detail data information can be provided but previous steps could be wrong, it will cost lots of resource and time to do re-collecting step. As for the questions asked after initial data processing, descriptive statistics could be provided and a series of actions or steps already been performed on the data, which could give the respondent a clear overview of the situation.

Based on all interview and content analysis findings, we propose some design implications for Q&A service helping for choosing statistical tests. The system should figure out question timing at the very beginning, since when people ask questions decide what information can be provided at this stage. According to timing information, the system should also give hint on what information should

be provided and help structure the information in a good manner. The system should also have the fabrication function to help users who do not want to provide their own data fabricate datasets and variables. It is better that system can alarm the questioner with every comment and encourage questioners to complement the lacking information. Lastly, tags information is quite important to locate questions within their expertise and provide answers with higher quality, so that the system should also automatically generate and allow experts to modify the question tag.

Appendix A

Informed consent form

Principal investigator:

Yue Hu
RWTH Aachen University
yue.hu1@rwth-aachen.de

Purpose: The goal of this project as a whole focuses on the improvement statistic analysis, with particular interest in how data analyst finds useful information during their study, how they ask questions and the handle problems. The study does not aim to evaluate your techniques or experiences. Rather, I am trying to learn more about statistical analysis and the data collected during this study will be used to understand statistical analysis workflows and strategies to cope with problems that arise during the process.

Procedure: Participation in this study involves a discourse to understand user's experience and background information, followed by a series of semi-structured questions. The experimenter will capture the screen and record the audio of the entire session. All information will be confidential. (See 'Confidentiality' below for details.)

Risks/Discomfort: Even though the study is expected to last no longer than one hour, you may become fatigued during the course of your participation in the study. Feel free to take as many breaks as necessary

during the study. There are no risks associated with participation in the study. Should completion of the task becomes distressing to you, it will be terminated

Confidentiality: All information collected during the study period will be kept strictly confidential. You will be identified only through identification numbers and background information you divulge in publications or reports. If you agree to join this study, please sign your name below.

I have read and understood the information on this form.

I have had the information on this form explained to me.

Participant:

Investigator:

Signature

Date

Signature

Date

Appendix B

Interview Protocol

The goal of this project as a whole focuses on the improvement statistic analysis, with particular interest in how data analyst find useful information during their study, how they ask questions and the handle problems. The study does not aim to evaluate your techniques or experiences. Rather, I am trying to learn more about statistical analysis and the data collected during this study will be used to understand statistical analysis procedure and strategies to cope with problems that arise during the process.

B.0.1 Interviewee Background

1. How long have you been doing statistical analysis(since the first time you know it)?
2. Have you take any statistical analysis course?
3. When you did your practical study later do you think this information you got in the course is helpful?
4. What's the topic of your Master thesis?
 - (a) Did you include statistical analysis part in it?
 - (b) How did you decide to do it?

B.0.2 Walkthrough a prior task

5. What's is your recently completed study included statistical analysis?
 - (a) When you design the experiment you already consider the statistical anslysis part?
 - (b) Which test method did you choose?
 - (c) How did you decide which test methods to use?
 - (d) Are you sure or confident about if you choose the right test methods?
6. Which tools did you use for this statistical analysis study?
 - (a) Some other tools you use for your statistical analysis?
 - (b) How did you know these tools?
 - (c) How did you learn to use this software? Is it user-friendly or have you met some problems when using it?
 - (d) Are you sure or confident about if you choose the right test methods?
7. After finish statistical analysis part ,what result did you get?
 - (a) Is the result significant?
 - (b) Have you ever met the situation that the result is not good?such as the p value or effect size is not good as expectation?
8. What about the report you write of this statistical analysis study. What did you include,what is the structure?
 - (a) Normally what is the structure of your statistical analysis report? Will you include detail information of your statistical analysis study such as standard derivative means?
 - (b) Did you asked someone to check the study process and the report?
9. What do you think the most difficult part of this statistical analysis?

10. Did you met some problems during the analysis?
11. How did you solve this problem?
12. If you had a magic wand ,what would you change to improve this statistical analysis process?
13. How long did you take for this statistical analysis study?
 - (a) How long do you think you spend on information gathering and finding? such as googling, finding useful papers, finding the proper tools ...
 - (b) Normally how long will a statistical analysis take in you previous experience?

B.0.3 Summary question

14. How do you think of statistical analysis and why do you have that feeling?
15. Normally you work alone on statistical analysis part or is a team work?
16. How will you share info with each other?How you discuss with others?
17. Normally how you find information which is useful for your statistical analysis? which do you think is most useful or effective?
18. Did you go to QA sites such as Stackoverflow to ask questions or find the useful answer about statistical analysis?
 - (a) Which website you used?
 - (b) Can you remember one scenario what questions you asked or what answers you are trying to find?
 - (c) Did you get the answer you want?
 - (d) How did you judge the answer is right or useful?

19. Do you think the useful material is easy to find for statistical analysis in HCI field (on google/on internet...)?
20. What advise would you give to the person who just start statistical analysis in HCI field?

Bibliography

- [1] Lada A Adamic, Jun Zhang, Eytan Bakshy, and Mark S Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, pages 665–674. ACM, 2008.
- [2] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding high-quality content in social media. In *Proceedings of the 2008 international conference on web search and data mining*, pages 183–194. ACM, 2008.
- [3] Mats Alvesson and Stanley Deetz. *Doing critical management research*. Sage, 2000.
- [4] David R Anderson, Dennis J Sweeney, Thomas A Williams, Jeffrey D Camm, and James J Cochran. *Statistics for business & economics*. Nelson Education, 2016.
- [5] Krisztian Balog, Leif Azzopardi, and Maarten De Rijke. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50. ACM, 2006.
- [6] RS Barbour. Theorizing in qualitative data analysis. dans rs barbour (éd.), *introducing qualitative research. a student’s guide to the craft of doing qualitative research* (pp. 232-254), 2008.
- [7] Nadia Boukhelifa, Marc-Emmanuel Perrin, Samuel Huron, and James Eagan. How data workers cope with uncertainty: A task characterisation study. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3645–3656. ACM, 2017.

-
- [8] Leanne Bowler, Jung Sun Oh, Daqing He, Eleanor Mattern, and Wei Jeng. Eating disorder questions in yahoo! answers: Information, conversation, or reflection? *Proceedings of the American Society for Information Science and Technology*, 49(1):1–11, 2012.
- [9] Jana Bradley. Methodological issues and practices in qualitative research. *The Library Quarterly*, 63(4):431–449, 1993.
- [10] John D Brewer. *Ethnography* open university press, 2000.
- [11] Alan Bryman. *Social research methods*. Oxford university press, 2016.
- [12] Paul Cairns. Hci... not as it should be: inferential statistics in hci research. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 1*, pages 195–201. British Computer Society, 2007.
- [13] Kathy Charmaz. *Constructing grounded theory: A practical guide through qualitative analysis*. sage, 2006.
- [14] Kathy Charmaz. *Constructing grounded theory: A practical guide through qualitative analysis*. sage, 2006.
- [15] Kathy Charmaz. Grounded theory as an emergent method. *Handbook of emergent methods*, 155:172, 2008.
- [16] Erik Choi, Vanessa Kitzie, and Chirag Shah. Developing a typology of online q&a models and recommending the right model for each question type. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–4, 2012.
- [17] Erik Choi, Vanessa Kitzie, and Chirag Shah. Investigating motivations and expectations of asking a question in social q&a. *First Monday*, 19(3), 2014.
- [18] Jacob Cohen. The world is round (p. 05). *American Psychologist*, 49(12):997–1003, 1994.
- [19] John W Creswel. *Research design: Qualitative, quantitative, and mixed methods approaches*. Los angeles: University of Nebraska–Lincoln, 2009.

- [20] Michael Crotty. *The foundations of social research: Meaning and perspective in the research process*. Sage, 1998.
- [21] Roger Purves David Freedman, Robert Pisani. *Statistics(4th ed.)*. W. W. Norton Company, Feb 13, 2007.
- [22] Zoltan Dienes. Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6(3):274–290, 2011.
- [23] Ruth E Fassinger. Paradigms, praxis, problems, and promise: Grounded theory in counseling psychology research. *Journal of counseling psychology*, 52(2):156, 2005.
- [24] Wei Feng and Jianyong Wang. Incorporating heterogeneous information for personalized tag recommendation in social tagging systems. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1276–1284. ACM, 2012.
- [25] Uwe Flick. Qualitative research designs. *Designing qualitative research*, Sage Publications, 2007.
- [26] Institute for Digital Research and Education UCLA. What statistical analysis should I use? <http://stats.idre.ucla.edu/other/mult-pkg/whatstat/>, 2015.
- [27] from Statistics Learning Center (Dr. Nic). Choosing which statistical test to use.
- [28] Rich Gazan. Social q&a. *Journal of the American Society for Information Science and Technology*, 62(12):2301–2312, 2011.
- [29] Barney G Glaser and Anselm L Strauss. *Discovery of grounded theory: Strategies for qualitative research*. Routledge, 2017.
- [30] Wayne D Gray and Marilyn C Salzman. Damaged merchandise? a review of experiments that compare usability evaluation methods. *Human-computer interaction*, 13(3):203–261, 1998.

-
- [31] Zoltan Gyongyi, Georgia Koutrika, Jan Pedersen, and Hector Garcia-Molina. Questioning yahoo! answers. Technical report, Stanford InfoLab, 2007.
- [32] F Maxwell Harper, Daphne Raban, Sheizaf Rafaeli, and Joseph A Konstan. Predictors of answer quality in online q&a sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 865–874. ACM, 2008.
- [33] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.
- [34] Kateryna Ignatova, Cigdem Toprak, Delphine Bernhard, and Iryna Gurevych. Annotating question types in social q&a sites. In *Tagungsband des GSCL Symposiums 'Sprachtechnologie und eHumanities*, pages 44–49. Citeseer, 2009.
- [35] James B Jones. Research fundamentals: Statistical considerations in research design: A simple person's approach. *Academic Emergency Medicine*, 7(2):194–199, 2000.
- [36] Maurits Kaptein and Judy Robertson. Rethinking statistical analysis methods for chi. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1105–1114. ACM, 2012.
- [37] George J Klir and Bo Yuan. Fuzzy sets and fuzzy logic: theory and applications. *Upper Saddle River*, page 563, 1995.
- [38] Samuel Kotz, Horman Lloyd Johnson, and Campbell B Read. Encyclopedia of statistical sciences. Technical report, 1982.
- [39] Steinar Kvale. *InterViews: an introduction to qualitative research interviewing*. Sage, 1996.
- [40] Jin Ha Lee, J Stephen Downie, and Sally Jo Cunningham. Challenges in cross-cultural/multilingual music information seeking. 2005.

- [41] Yvonna S Lincoln. Emerging criteria for quality in qualitative and interpretive research. *Qualitative inquiry*, 1(3):275–289, 1995.
- [42] Yvonna S Lincoln and Norman K Denzin. *The handbook of qualitative research*. Sage, 2000.
- [43] Thomas D Lonneman Doroff. Supervision in applied counseling settings: a socially constructed grounded theory. 2012.
- [44] Mangiafico. Summary and analysis of extension program evaluation in r. <http://rcompanion.org/handbook/> ., S.S. 2016.
- [45] Marius Marusteri and Vladimir Bacarea. Comparing groups for statistical differences: how to choose the right statistical test? *Biochemia medica: Biochemia medica*, 20(1):15–32, 2010.
- [46] Jennifer Mason. *Qualitative researching*. Sage, 2017.
- [47] JH McDonald. Handbook of biological statistics., 3rd edn.(sparky house publishing: Baltimore, md.). 2014.
- [48] Victor Minichiello, Rosalie Aroni, Eric Timewell, Loris Alexander, et al. In-depth interviewing: Principles, techniques, 1995.
- [49] Audris Mockus and James D Herbsleb. Expertise browser: a quantitative approach to identifying expertise. In *Proceedings of the 24th International Conference on Software Engineering. ICSE 2002*, pages 503–512. IEEE, 2002.
- [50] Barun K Nayak and Avijit Hazra. How to choose the right statistical test? *Indian journal of ophthalmology*, 59(2):85, 2011.
- [51] Hebatallah Nour-Eldein. Statistical methods and errors in family medicine articles between 2010 and 2014-suez canal university, egypt: A cross-sectional study. *Journal of family medicine and primary care*, 5(1):24, 2016.
- [52] Aamir Omair et al. Understanding the process of statistical methods for effective data analysis. *Journal of Health Specialties*, 2(3):100, 2014.

-
- [53] Alex T Pang, Craig M Wittenbrink, and Suresh K Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, 1997.
- [54] Swati Patel, VD Naik, and Prakash Patel. Use of statistical methods and complexity of data analysis in recent research publications in basic medical sciences. *Community Med*, 5(2):253–256, 2014.
- [55] Michael Quinn Patton. *Qualitative evaluation and research methods*. SAGE Publications, inc, 1990.
- [56] Steffen Rendle, Leandro Balby Marinho, Alexandros Nanopoulos, and Lars Schmidt-Thieme. Learning optimal ranking with tensor factorization for tag recommendation. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 727–736. ACM, 2009.
- [57] Lyn Richards. *Handling qualitative data: A practical guide*. Sage, 2014.
- [58] Judy Robertson and Maurits Kaptein. *Modern statistical methods for HCI*. Springer, 2016.
- [59] Helen B Schwartzman. *Ethnography in organizations*, volume 27. Sage, 1993.
- [60] Chirag Shah, Sanghee Oh, and Jung Sun Oh. Research agenda for social q&a. *Library & Information Science Research*, 31(4):205–209, 2009.
- [61] Tejinder Singh. Research methodology simplified: Every clinician a researcher. *Indian Journal of Pharmacology*, 43(2):224–224, 2011.
- [62] Ruud Wetzels, Dora Matzke, Michael D Lee, Jeffrey N Rouder, Geoffrey J Iverson, and Eric-Jan Wagenmakers. Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, 6(3):291–298, 2011.
- [63] Wikipedia contributors. Summary statistics - Wikipedia, the free encyclopedia.
- [64] Wikipedia contributors. Qa software Wikipedia, the free encyclopedia, 2019, July 26. [Retrieved 14:03, August 2, 2019].

- [65] Yan Zhang. Contextualizing consumer health information searching: an analysis of questions in a social q&a community. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 210–219. ACM, 2010.

Index

comments.....	36
content analysis.....	20
criteria.....	19
qualitative coding.....	20
ResearchGate.....	18
statistical methods resources.....	6
abbrv.....	<i>see</i> abbreviation
ambiguous sentence.....	59–60
ask question.....	36
assume.....	54, 59, 63
assumption.....	43, 44
books.....	31–32
chaotic structure.....	56
charts.....	57–59, 61
common methods.....	26
common methods.....	27, 71
communication.....	33
data privacy.....	76
data processing.....	68
dataset.....	42
dependent variables.....	47
design implication.....	79
different options.....	55, 63
domains of Q&A websites.....	8
edit function.....	18, 56
example.....	40
experiment design.....	46
experimental procedure.....	48
fabricate dataset.....	54
Fabricated variables.....	47
focus coding.....	17
grounded theory.....	15

hard finding experts	75
hypothesis	48
inaccurate phrase	59
inappropriate statistical methods	5
independent variables	47
information credibility	74
initial coding	16
interview method	11–12
MAXQDA	16
memo writing	16
missing information	52–55
modle's interpretability	72
negative attitudes	27, 31, 36
new methods	28, 32
One question asking for code script	41
papers	32
participants number	13
process methods	62
processing method	53
Purposeful sampling	12
Qualitative approach	10–11
qualitative coding	16–17
question timing	64–68, 78
questioners motivations	7
result interpretation	39
sample size	48
search similar question	36
semi-structured interview	12, 14
software	49
Stack Exchange	18
study process	64
summary data	42
tag	18, 61
tags	50, 77
taxonomy of questions	7
themes	23, 37
time limitation	73
uncertainty	23–25, 39, 69
uncertainty in statistical analysis	1, 5, 9
unclear information	55
upvote number	36

within/between-group design 48
wrong methods 25

