University of Zurich UZH

Thesis April 23, 2015

CrowdSA: A crowdsourcing platform to extract and verify the correct usage of statistics in research publications

Mattia Amato
of Zürich ZH, Switzerland

Student-ID: 10-919-686
mattia.amato@uzh.ch

Advisor: **Patrick de Boer, Michael Feldman**

Prof. Abraham Bernstein, PhD
Institut für Informatik
Universität Zürich
http://www.ifi.uzh.ch/ddis

CrowdSA: A crowdsourcing platform to extract and verify the correct usage of statistics in research publications

Mattia Amato

April 2015

1 Abstract

This thesis aims to offer a new kind of approach to solve the statistical flaws in research by helping people to efficiently extract information from research publications. The statistical flaws is a problem which afflicts many research fields by creating wrong discoveries. This issue has also an impact on daily life since the discoveries carried out from scientific publications are used everyday in different occasions, e.g., in the medical field. The solution proposed to reduce this problem is to create a new crowdsourcing platform, CrowdSA, which outsources the complex work of the reviewers to the crowd. This system is able to extract from any kind of publications several statistical methods and validate them. The extraction as well as the validation are performed by distributing different questions to the crowd and collecting their answers.

2 Related work

The available literature offers a wide overview over the existing crowdsourcing systems and the problem of statistical flaws in research. A brief introduction about the crowdsourcing topic is given, as well as an analysis of the statistical flaws problem which is discussed in many scientific publications. In a second section, a research was performed on the existing crowdsourcing platforms in order to identify similar systems.

2.1 Crowdsourcing

Crowdsourcing is a technique which allows people to solve complex problems in a relative small amount of time with very low costs. Outsource or dislocate the activities are well known techniques which are widely used today in the economical field to exponentially increase the output with only a linear increase of the costs. The main idea behind crowdsourcing is based on the assumption that different people working on a task create equals or even better and more complete results than a single worker solving the same problem on its own. This phenomenon is also called *the wisdom of the crowds* [1].

One of the particularities of crowdsourcing systems is how to organize the crowd. Several studies tried to identify an efficient structure to manage workers who are not bound to each other, e.g., as in a company structure. One of the solutions proposed is to differentiate the workers based on the results they achieve. This will give more responsibility to some workers than to others [2]. The qualification system, based on the achieved results, is used in several crowdsourcing platforms, e.g., Wikipedia or MTurk. The fundamental reason to differentiate between the crowd workers in this way is that if one contributes to solve a problem with engagement and delivers good solutions, then this fact can be used as a guarantee for the future tasks [3]. Others important questions widely discussed in the crowdsourcing field are: how to bind the workers to a marketplace which offers the tasks? What are the incentives which best stimulates the workers to give truly and complete answers?

Binding people to web based platforms is a problem that is faced every day in the world wide web. The most common solutions to this issue are: aggressive marketing techniques, the use of monetary incentives as well as create network effects [4]. Nowadays the diffusion of technology makes the network effect one of the strongest forces to bind people to a web platform. The incentives for a worker, on the other hand, can be summarized in two main categories, physical, e.g., monetary rewards, or psychological, e.g., comparing the work done with the results of other workers or being aware of contributing to an important cause [5].

The opportunities as well as the needs satisfied by the crowdsourcing systems are many: from finding solutions to complex problems, to gathering more information about specific topics[6]. The externalization of the tasks to the crowd gives the users the possibility to quickly reach a high number of participants paying only few cents per answer. An example of such a system is MTurk[1]: this platform allows to create different tasks which are available to crowd workers. All of them, or in some cases only the workers who meet specific requirements, are able to give a solution through the platform and earn some money.

2.2 Statistical flaws in research

An important problem in research papers is the correct usage of statistical means as well as the method used to collect the data for an experiment. In order to ensure the correctness of these publications, there are several reviewers who constantly proof research papers and find basic statistical errors. The misinterpretation of the results, or the errors in using statistical means over wrongly collected datasets, has a strong impact on the discoveries described in the publications. This kind of mistakes may have different influences on daily life, e.g., a research may consider a medicament, after several experiments, adequate to treat a certain disease. If the results of the experiments are misinterpreted, it would be possible that the medicament does not have the desired effect on the patient.

Several researchers analysed this problem and reported a high number of errors in the usage of statistics. One of these researches [7] analysed 139 articles from the *Korean Journal of Pain*[2] published between 2004 and 2008 and found out that only 20.9% of the articles were free from statistical errors. The most common error identified by this research was '*no statistics used even though statistical methods were required*' (40.6%) [7]. This kind of researches proves that it is necessary to assess the statistics before publishing a scientific research.

The statistical flaws is present in almost every research field [8]: several attempts have been made in order to solve it, some with positive results, but a final solution has not been defined

[1] http://www.mturk.com/

[2] http://koreamed.org/

yet. An interesting point of view regarding this problem is described in [9]. This research analysed 51 publications from 2005 to 2009 containing 66 empirical studies. The results identified a common pattern in using self-developed questionnaires without providing items or statistical validations [9]. This kind of errors are very common [9] and may also introduce, in the worst case, important miscalculations in the statistics.

An attempt to solve this issue, e.g., in the biological research field, was conducted by Nature - a prominent interdisciplinary scientific journal - which discussed the problem in different articles and adopted some policies to solve the problem [8]. One of them consists in insisting that papers containing figures with error bars have to contain some information that describes what the error bars actually represent. Introducing these policies did not stop the problem [8].

Nowadays it is still possible to find in this and in other similar journals different researches containing exactly the same problems which were supposed to be solved [8]. Another solution proposed by David L. Vaux, a professor of cell biology from the Melbourne University, is to follow the lead of the *Journal of Cell Biology* and make a final check of all figures in the accepted papers before the publication, as well as to refuse to publish papers that contain fundamental errors, and readily publish corrections of already published papers [8]. In order to achieve this, experts in the statistical field should review the papers before their publication and inform the researchers about the errors contained in their documents [8].

We believe that the complex task of reviewing scientific researches can be supported by a crowdsourcing platform. This system should reduce the complexity of the work by dividing the main problem into multiple subproblems which can be quickly and easily solved also by non-expert of the field.

2.3 Similar Systems

The main goal of this thesis is to implement a crowdsourcing platform that helps the reviewing process of scientific researches. This process, as described before, is a highly complex task which can only be executed by workers who have at least some basic knowledge about statistics.

Nowadays it is possible to easily find many different crowdsourcing platforms in internet which support any user in solving complex tasks. Currently, the most known crowdsourcing based platforms are: Wikipedia[3], Linux[4], Yahoo! Answers[5] and Mechanical Turk based systems [6]. In this last category there are several different systems such as Amazon Mechanical Turk (MTurk), CrowdFlower, FlashTeams or CrowdWeaver [6]. The goals of these crowdsourcing systems are different from each other. Wikipedia's goal, e.g., is to constantly increase the quality and the quantity of the knowledge stored in the system whereas Mechanical Turk based systems are used to distribute complex tasks in order to solve them in a short amount of time and with relative small costs. The quality of the output produced, which should be as higher as possible, is what all these systems have in common.

It seems that today there is no similar system such as the Crowd Statistical Assessment (CrowdSA) platform. This particular crowdsourcing platform binds each question to a specific paper, stored in the server as a PDF file. This peculiarity differentiates CrowdSA from all

[3] http://www.wikipedia.com/

[4] http://www.linuxfoundation.org/

[5] http://answers.yahoo.com/

the other existing crowdsourcing systems.

There are different peer-reviewer platforms in internet which allow to discuss a research as well as to distribute the paper to the crowd and waiting for a final feedback. The difference between these reviewing systems is that CrowdSA is a transparent crowdsourcing platform supporting the reviewers work with integrated online tools.

3 Analysis of research publications

Before starting to build a solution to solve the problem of assessing the statistical means in scientific publications, a research was performed to gain further information about what really needs to be extracted from the papers and how this information can be used to validate a publication. Several papers from the CHI conference[6] and the BMC medical conference[7] were analysed in order to identify a common pattern in the usage of statistical means.

3.1 What needs to be extracted?

There are three main elements which need to be extracted to verify whether the statistical methods are correctly used over a specific research: datasets, statistical methods and the relation between the previous two elements.

A dataset is composed by all the relevant information about the population participating in a specific test, e.g: range of ages, size of population and nationality. The analysis of this central element also showed a multitude of different ways to define it. Some of them are described in the same document, other are identified in external papers and some are only accessible through an URL. It is also possible that a paper contains multiple datasets or just refers to a collection which is not presented in details to the reader.

On the other hand, to verify if a particular statistical method is used in a paper, a database containing different definitions of statistical methods is used. Each element in this database also includes a list of assumptions which need to be validated in order to correctly use the method. An assumption is, for instance, *Normality* or *Linearity* and can be inquired with several tests. The *Normality* can be tested using, e.g., the *D'Agostino K-Squared test* or the *Shapiro-Wilk test*.

If the statistical methods can be identified at runtime by a simple text match function, this is not true for the identification of the datasets and the relations. These two elements are too complex and differ, in their structure and definition, from paper to paper. For this reason human work is needed to successfully complete the extraction task.

4 Experiment setup

An experiment was executed in order to better estimate the budget and time needed by the platform to complete the evaluation of five different BMC publications. The results of this experiment were compared, in a second phase, to the official review written by experts of the sector.

First a brief introduction to a library called PPLib [10], developed by the University of Zürich[8] (under review), is given. This library was integrated into the CrowdSA application to simulate the different patterns and collect the

[6] http://chi2015.acm.org/

[7] http://www.biomedcentral.com/

[8] http://www.uzh.ch/

necessary information to measure the efficiency of each process. After which a list of hypotheses to increase the productivity of the crowd workers and to ensure the high quality of the output are presented.

The last section concerns the general workflow of the CrowdSA application, from finding the statistical methods in the paper to the extraction of the datasets and relations as well as the final evaluation of the publication.

4.1 PPLib

PPLib (pronounced "People Lib") is a library written in Scala that allows to apply several recombinations of processes to existing crowdsourcing systems and evaluate which of these recombinations is the most efficient in terms of budget, time and output quality [10]. PPLib has already been used successfully to translate texts from German to English and to shorten long texts as described in [10]. This library is comparable to Turkit [11] where the developer create a process which asks questions to crowd workers and acts upon their answers. One particularity of PPLib is that it is built to be an independent Human Computation Platform [10]. Therefore PPLib can be used as an intermediary to communicate with several crowdsourcing platforms like CrowdFlower[9], MTurk[10] or as in this example, CrowdSA.

4.1.1 Recombinations

As mentioned, PPLib recombines different processes and evaluates the best alternative. The PPLib Process Repository (PPR) was extended to match the general workflow of the CrowdSA platform. The PPR is organized in a taxonomic structure and allows to store and execute the processes applying different patterns [10]. As described in [10], the structure of the PPR is inspired by the *Process Handbook* [12] and the *collective intelligence genome* [13].

It is important to notice that the PPR contains the top-level genomes suggested in the WHAT dimension of their ontology: CREATE and DECIDE [10]. All the processes developed for CrowdSA are based on these two genomes too. CREATE is used to collect information from crowd workers, whereas DECIDE is used to select one or more collected information [10], e.g., through a voting process.

The recombinations are variation of the processes defined in the PPR. The processes can have variations, for example, in the number of crowd workers who are allowed to work on a single process or in the confidentiality interval a process has to use. All these variations are automatically generated by the PPLib Recombinator generator. The processes that can be used in CrowdSA are described in the next sections.

4.2 Collect-Decide

Collect-Decide is a two step pattern. As the name suggests the *collect phase* brings together the information that is analysed in the second step where a decision has to be carried out. This pattern is divided in a CREATE and a DECIDE process. The *collect phase* is supported by two different processes: *Collection with sigma pruning*, or the more generic variant called *Collection*. The first process collects the information and automatically prunes the data, e.g., by text length or distinct values, whereas the second process collects the information without any particular restriction. These two processes can be executed

[9] http://www.crowdflower.com/

[10] http://www.mturk.com/

with different variations, e.g., by setting for each instance a different number of required crowd workers. For the *decision phase* a normal *Contest* process is used.

The other available processes for this step are *Beat-By-K* [14] and the *statistical reduction voting process*. The *Beat-By-K* voting process creates voting questions until one answer reaches K more votes than all the others alternatives. On the other hand a *statistical reduction voting process* creates single voting questions until a configurable confidence parameter is reached.

4.3 Iterative Refinement

This particular pattern is, as the Collect-Decide pattern, based on a CREATE and a DECIDE process. The information collected in the first step is used in the next question where the worker is asked to refine it. This refinement process can last for several rounds until it converges to an answer which is supposed to represent the truth. If the refined information differs from the previous one, a voting question is created to choose which information to refine next. This process converges when the information does not change.

4.4 Hypotheses

Several hypotheses have been postulated to increase the efficiency of the platform.

H1: Freedom of choice for the crowd workers

Since the particularity of CrowdSA is that all the questions are related to a scientific research displayed as a PDF file, one of the central hypothesis is that the crowd workers perform better if they can choose the research topic to analyse without any particular restriction. This is due to the personal interests of the workers who may vary from one to the other. Working on an interesting topic should automatically increase the time a person wants to work on it and the will to respond truthfully. Likewise, it is also important for a crowd worker to be able to choose among a list of available questions, since answering mandatory questions decreases the performance of the single worker. For this reason the CrowdSA platform offers five different layouts to the users. These layouts differ in the way crowd workers get the questions: on one side a random selection is performed whereas on the other side the user can freely choose them. In this last case the questions can be filtered by reward, type or research topic.

H2: Ranking system increases the productivity of the worker and the quality of the answers

Another important hypothesis postulated was the increase of the productivity as well as the quality by displaying the users their ranking position. This phenomenon, which is described in several researches about the gamification topic, seems to stimulate the workers to do a better job. Being aware of the personal position into the whole system creates a challenge between the users that can be won only by delivering better solutions and answering more questions.

H3: Tools which support the crowd workers will increase the overall performance

Supplying the right tools to work with,will increase the productivity which also affects the overall performance of the workers. For this reason an embedded PDF viewer was integrated in the CrowdSA application and a highlighting function was developed to support the work process. Highlighting the relevant terms and being able to jump between the searched terms is a fundamental functionality which needs to be present to reduce the complexity of the work.

4.5 General Workflow

CrowdSA is a platform divided in a client and in a server application. These two applications interact together as shown in figure 1. When a user wants crowd workers to validate a scientific publication, the client application will create all the questions in an automatic way and delivers a feedback with a final evaluation of the paper. The research to be evaluated goes through different processes.

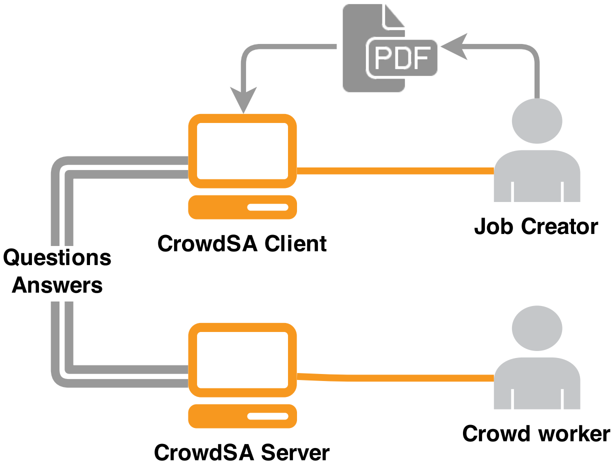


Figure 1: Overview of the CrowdSA application. The client creates the questions that are answered by the crowd workers.

First of all, it is analysed in order to find the matches with the statistical methods defined in the database. In a second step, the discovery phase, an extraction process is started for each identified statistical method. Since each statistical method is assumed to be related to a dataset, this step asks the crowd workers to identify it. The third step consists of ensuring that all the assumptions related to a specific statistical method are respected. This last step is entirely composed by binary questions.

At the end, in a fourth step, a final feedback of the scientific research is generated by analysing the collected information from the previous processes.

4.5.1 Automatic match of statistical methods

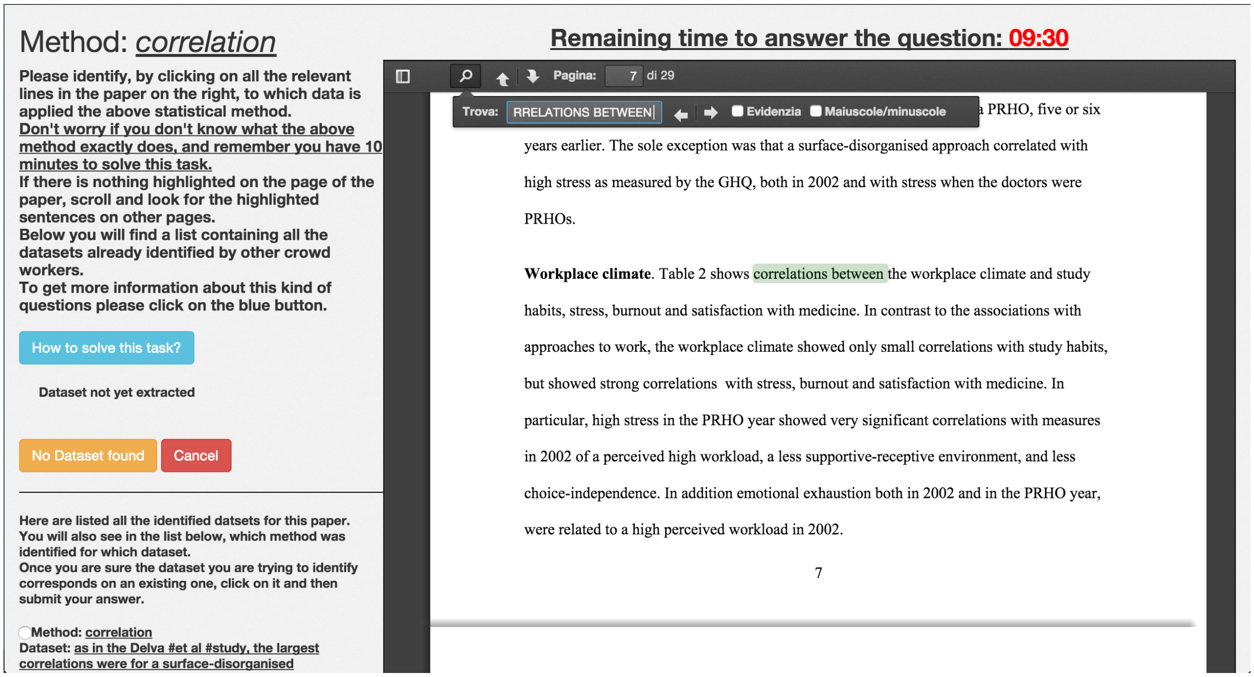
The easily extendible database contains 27 different statistical methods which can be matched to any paper. At first the matching process loads the research as a PDF file into memory and extracts all the text contained in it, then searches for each statistical method defined in the database if a match occurs.

Since PDF is a standard developed by Adobe Systems[11] in 1993, using libraries that are not developed by this company to extract elements from the pages is quite a complex task which may create distortions in images, tables or text structure.

A library called PDFBox[12] was used to extract the text. This library is not officially supported by Adobe Systems and introduces some minor distortions but also allows to manage PDF files by extracting text, images, tables and adding elements such as geometric figures and annotations to the pages without losing the standard format defined by Adobe Systems. After the matches

[11] http://www.adobe.com/

[12] http://pdfbox.apache.org/



Method: *correlation*

Remaining time to answer the question: 09:30

Please identify, by clicking on all the relevant lines in the paper on the right, to which data is applied the above statistical method.
Don't worry if you don't know what the above method exactly does, and remember you have 10 minutes to solve this task.
If there is nothing highlighted on the page of the paper, scroll and look for the highlighted sentences on other pages.
Below you will find a list containing all the datasets already identified by other crowd workers.
To get more information about this kind of questions please click on the blue button.

Dataset not yet extracted

Here are listed all the identified datsets for this paper.
You will also see in the list below, which method was identified for which dataset.
Once you are sure the dataset you are trying to identify corresponds on an existing one, click on it and then submit your answer.

○ Method: correlation
Dataset: as in the Delva #et al #study, the largest correlations were for a surface-disorganised

RRELATIONS BETWEEN

a PRHO, five or six years earlier. The sole exception was that a surface-disorganised approach correlated with high stress as measured by the GHQ, both in 2002 and with stress when the doctors were PRHOs.

Workplace climate. Table 2 shows correlations between the workplace climate and study habits, stress, burnout and satisfaction with medicine. In contrast to the associations with approaches to work, the workplace climate showed only small correlations with study habits, but showed strong correlations with stress, burnout and satisfaction with medicine. In particular, high stress in the PRHO year showed very significant correlations with measures in 2002 of a perceived high workload, a less supportive-receptive environment, and less choice-independence. In addition emotional exhaustion both in 2002 and in the PRHO year, were related to a high perceived workload in 2002.

7

Figure 2: Viewer page with question, embedded PDF reader and highlighted match

are identified in the text, for each of these elements a discovery process is started.

An attempt to automatically extract statistical methods was performed in [15]. It seems to be possible to extract most of the terms using only simple grammatical patterns with the support of NLP and machine learning techniques. The main goal of the above-mentioned research was to identify the statistical methods used especially in the biomedical field, in order to be able to build glossaries, ontologies and specialist lexicons [15]. The results collected in their experiment showed, for a rule-base approach, a prevedibile high recall (100%) and a precision of 85.40% [15]. The machine learning approach had a recall of 75% and a precision of 81.9% showing that there are still some possibilities of improvement in this field [15].

4.5.2 Discovery step

The discovery step (figure 3) is a central element of the platform. In this step the client application creates the questions for the crowd workers with the goal to successfully identify the dataset related to a specific statistical method in the paper. An example of such a question is: *"Please identify the dataset of the statistical method: ANOVA highlighted in the paper"* where the *ANOVA* method is highlighted and the worker has to identify to which variables and attribute it is related.

These questions, which are generated by a CREATE process in the client, are sent to the server application. Once a crowd worker accept to answer a particular question of this step, the viewer page is loaded (Figure 2). In the

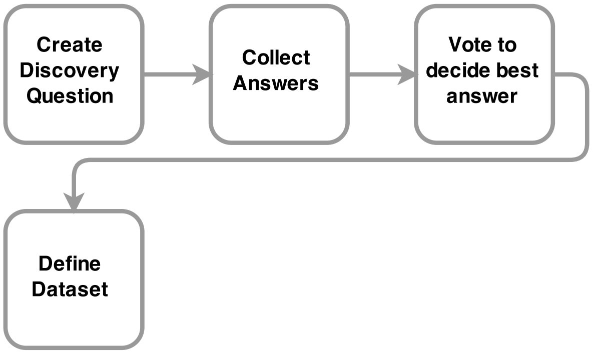


Figure 3: Discovery step

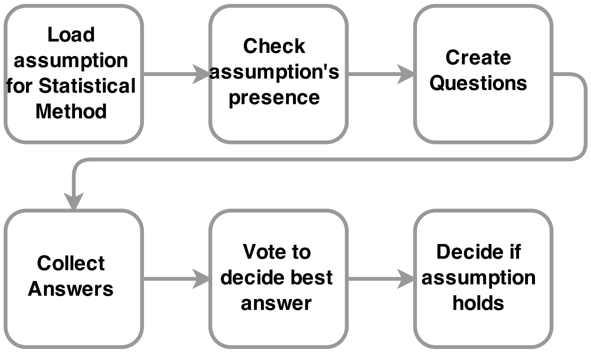

Load assumption for Statistical Method
Check assumption's presence
Create Questions
Collect Answers
Vote to decide best answer
Decide if assumption holds

Figure 4: Assumption step

viewer page, on the left side, the questions and the instructions are displayed defining what exactly the crowd worker has to do to successfully solve the task. On the right side, an embedded - JavaScript based - PDF viewer is loaded, showing the document with the relative highlights related to the question.

The PDFjs library[13] allowed to integrate a PDF viewer to this page. The functionalities offered by this library are several and support the work of the crowd workers.

Once all the answers to the discovery questions are collected by the client application, a DECIDE process starts. The latter, e.g., in a *Collect-Decide* phase with *Beat-By-K* voting process, analyses the answers of different crowd workers and compares them. If the answers are different from each other a voting question is created with the purpose to decide which alternative is the most accurate. At the end of the discovery step, a dataset is identified and created in the server application. Future questions will refer to this dataset in order to help the reviewers to validate the assumptions.

4.5.3 Assumptions step

The assumptions step (figure 4) is executed after a dataset has been created on the server and repeated every time a new dataset is identified. The goal of this step is to verify whether all the assumptions hold for a specific statistical method.

For instance, the *ANOVA* method needs the assumptions of *Normality*, *Homogeneity of variances*, *Constant Variance* as well as the *Independence* of the dataset to be respected. If one of these assumptions is not respected then the *ANOVA* method cannot be used and some errors may occur in the results.

In a first phase all the needed assumptions for a statistical method are loaded from the database. For each of these assumptions there are several tests which can be used to compute if it holds or not. The *Normality*, for example, can be tested with the *D'Agostino K-Squared* method or the *Shapiro-Wilk test*. These can be used to validate the assumptions. For this reason a list of possible tests is also defined in the database.

Each of these tests contains a list of possible

[13] http://mozilla.github.io/pdf.js/

matches that may help to identify their presence in the research. Because of this, a research through the whole publication needs to be performed in this step too.

In other words: first a list of tests, which can validate an assumption, are loaded into memory, then the presence of this tests is searched into the publication and if a match occurs, a question such as *"Is the dataset highlighted in the paper tested for Normality using the D'Agostino K-Squared test?"* is generated. On the other hand, if no match occurs, then a general question is created, for example *"Is the dataset highlighted in the paper tested for Normality?"*.

After this CREATE process, where the system collects the information about the used assumptions, a DECIDE process is started only if the workers gave different answers. This step depends on the process that has to be executed in the current variant., e.g., a normal *Contest* process or a *Beat-By-K* voting process. At the end of the assumptions step, for each needed assumption there is a converged answer which is supposed to represent the truth. This answer will be used in the next step to evaluate the research publication.

4.5.4 Evaluation step

The last step of the client application is to give a final evaluation based on the converged answers collected during the two previous steps.

PPLib allows to evaluate the different recombinations executed by storing the time needed to complete them, the cost of the processes and the final feedback created by the client application. For each dataset identified and statistical method, the converged answers are analysed.

If the answers related to an assumption are all negative, the dataset is evaluated as invalid. On the other hand, if at least one converged answer for a specific assumption shows that the dataset is correctly tested for it, the dataset is evaluated as valid.

5 Collected Data

A pilot experiment was run over two days with a total of 13 crowd workers to evaluate CrowdSA's output. This pilot test was based on five different papers from BMC Medicine[14] - *"a medical journal, publishing original research, commentaries and reviews that are either of significant interest to all areas of medicine and clinical practice, or provide key translational or clinical advances in a specific field"* - and used only the Collect-Decide process above described.

The estimation of time and number of workers necessary to run the experiment was too optimistic: the information collected during two days of test were not enough to complete all the five different processes. Since the estimation of the crowd worker's number as well as the number of hours needed to answer the questions of a single paper was not correct, only one out of five papers was solved in time.

The difficulty of foreseeing these variables for each paper was mainly caused by the high variation of the statistical methods matched in the different papers. Two out of five papers in the experiment had between three and six questions available for the first phase. The other three papers generated between 20 to 35 questions. This fact had important consequences on the time needed to answer all the questions.

The experiment was repeated with other six crowd workers and only one paper, which was

[14] http://www.biomedcentral.com/bmcmed/

chosen by the number of questions generated. It was successfully completed in about one hour.

In order to ensure that the crowd workers were not able to answer the voting questions containing part of their previous answers, a qualification system was also introduced. This kind of system limits the availability of any type of questions to a restricted portion of the crowd workers subscribed to the platform. It also increases the number of crowd workers needed but guarantees a correct voting process.

At the end of the experiments, the two completed papers were compared to the official reviews produced by experts. This allowed to estimate the correctness of the results. The analysis performed by the expert reviewers showed a high quantity of details which were not produced by CrowdSA, mainly because of the poor quantity of available questions in the database. The first paper, *"Outcomes of polytrauma patients with Diabetes Mellitus"* [16], was evaluated by six different crowd workers. For this particular paper, only three statistical methods were identified in the corpus. One of these methods was correctly considered to be a false positive match, since it was part of the list of the abbreviations used, whereas the other two methods, the *ANOVA* and *Chi-Square test*, were correctly identified and related to the same dataset.

All the assumptions needed by the *Chi-Square* method were validated by the crowd workers, suggesting that the dataset is correctly tested for each of them. On the other hand, the *ANOVA* method did not pass all the tests. The crowd workers identified that the dataset used to calculate the *ANOVA* method was not tested for *Normality*. Since *Normality* is an assumption which needs to hold in order to be able to use the *ANOVA* method, the platform evaluated the dataset as invalid for this statistical method.

The official report of the reviewer Daniel Denis, pointed out different observations which were not clear enough in the statistics of the publication, e.g., '*the authors should emphasize effect size and magnitudes instead of focusing too much on statistical significance*' or '*the p-values are not consistently reported. In some places p = 0.005 and in other p <.001*' . None of these points were identified by CrowdSA. The difference between the result of the platform and the one of the expert reviewer may be attributed to the Collect-Decide process, which collects, for the moment, only coarse information instead of asking questions regarding details that only workers with a strong statistic background would be able to answer correctly.

The second paper, *"Additional Saturday rehabilitation improves functional independence and quality of life and reduces length of stay: a randomized controlled trial"* [17], was analysed by 13 different crowd workers. Only two *ANCOVA* methods were identified on the paper. This allowed to keep a low number of questions. The paper was also solved faster compared to the others, indicating a preference for the crowd workers to answer to papers with a small amount of questions available. The crowd workers identified the same dataset for the two recognized methods present in the publication. The result reconstructed from the data collected by the platform revealed the need of rewriting the statistics of the paper for the two matched methods.

The outcome showed that all the assumptions except one, the *Linearity*, were not correctly tested. If this evaluation is compared to the official review provided by Dale Needham, a different result is presented. The expert reviewer pointed out that the paper did not need to be seen by a statistician since there was no particular problems with the statistics of the publication.

The difference in the final evaluation may suggest that the approach used by CrowdSA to validate a statistical method is not correct. Collecting a higher number of details about the properties of the datasets and the statistical methods may be the right solution to better validate a dataset. Despite this fact, in the official review, there are listed several points which need further clarifications by the writer of the research, e.g., it is mentioned that '*It appears that the primary outcome was changed from the published protocol based on the sample size calculation described in the manuscript, i.e. the sample size calculation described in the protocol was based on length of stay whereas the calculation described in the manuscript also took into account functional independence as a primary outcome*' or '*the authors state the goal enrollment was 968 patients, but according to Figure 1, 996 patients were randomized*'. All these minor problems seem not to have an impact on the statistics of the paper. In order to assess these small details, the introduction of the possibility to let the crowd workers analyse the statistical methods as well as the datasets with their own words and discuss them in a dedicated forum may create better results than the actual one.

5.1 Workers behaviour

Each of the recruited workers successfully created an account on the platform and was able to answer the available questions. The workers were distributed over different time zones and were paid 7 USD per hour of work. All of them actively participated to the experiment and in three different cases they asked for further clarifications via email or Skype.

The main problems reported were: the incompatibility of the web browser with the platform, even if was explicitly requested to only use Chrome, a problem related to the support of different time zones which was fixed successfully in time for the experiment and, in two different cases, the workers complained about the difficulty to use the platform since the instructions were not clear enough. After providing a real example, both crowd workers understood the structure of the questions and were able to continue their work.

5.2 Feedbacks from crowd workers

The interest of the crowd workers to participate in the experiment was high. Two different crowd workers complained in their feedbacks about the difficulty in solving the assigned task. The discovery phase seems not to be clear and easy enough to be solved with a minimum effort and, in some cases, the description of how to solve the task created some confusion.

Some crowd workers also reported a problem related to the highlighted terms. It was not clear enough why a particular element was highlighted and which role it had for a specific question. Another feedback from a crowd worker pointed out the need to extend the functionality to choose which lines contains the data for a specific method since there was some problems of highlighting mathematical expressions, e.g., p <.05.

All the feedbacks were submitted via email or communicated via Skype and were related to the discovery step. The instructions related to the questions of the assessment step were clear enough and nobody asked for further information.

6 Discussion

The results prove that CrowdSA is able to extract the statistical methods and verify the presence of the assumptions needed by asking few questions. This application can be used to support the work of the reviewers but cannot replace the entire work of an expert reviewer yet. Since the results only summarize the correct usage of some predefined statistical methods, for more complex papers it is necessary to extend the number of statistical methods present in the database as well as integrate a comment functionality to let the worker explain more precisely what exactly is tested, e.g., by asking a motivation for the answer given. For this purpose, the usage of a forum to discuss different answers can be a valid solution to increase the correctness of the answers and create a social network effect between the workers and the platform.

Another important point is how to present the questions to a worker. Since the discovery step has been considered too complex by several crowd workers, it is essential to rework the approach used to extract the dataset for each identified statistical method as well as to refine the questions asked to the crowd workers. It is also of primary importance to improve the way the instructions are presented for each question type and provide actual examples explaining again how to solve a question efficiently using the tools provided by the platform.

7 Conclusion

Nowadays, there are several crowdsourcing platforms which supports workers in solving repetitive and complex tasks. Many of these platforms, e.g., MTurk, CrowdFlower or FlashTeams, allow to create various tasks. On the other hand there are few platforms which are explicitly developed to support specific predefined processes like CrowdSA does. This specialized platform is still not ready to be used by non-statistically educated crowd workers and needs some improvements in the processes as well as in the instructions in order to increase its efficiency. The results of the experiment showed a lack of details which could have had an influence on the final evaluation computed by CrowdSA.

The lack of details could be attributed to the diversity of methods present in the database. Since the database contained only 27 different statistical methods, it can be a problem to correctly evaluate a publication. For this reasons the database needs to be extended and the questions contained in it simplified, so that they can be easily understood by workers who do not have any particular background in statistics.

The future steps to increase the efficiency and correctness of the platform could be summarized as follows: first of all, it is necessary to simplify the two steps process of the platform in order to reduce the complexity of the questions. There is also the need to model the possibility to collect more details about the statistical methods as well as datasets. The details requires to be deeply analysed to correctly evaluate if a dataset is tested or not for a certain assumption. Introducing a sort of forum or chat in order to discuss further details may improve the quality of the final evaluation since the crowd workers will be able to exchange and justify their own opinions.

Another important improvement could be the integration of NLP and machine learning techniques, as described in [15], which could generate better matches in the first phase and increase the validity of the final paper's evaluation.

All these facts suggest that CrowdSA is not ready to be used by crowd workers who do not have any particular statistical knowledge and needs further refinement in the questions asked as well as the given instructions.

Acknowledgements

I want to thank my advisors Patrick de Boer and Michael Feldman for the support during the development and experimentation of the CrowdSA application. I want also to thank my friends as well as my family who helped me with the experiment and reviewing the thesis.

References

[1] James Surowiecki. *The Wisdom of Crowds*, Doubleday, Anchor, 2004, 336 pages.

[2] Walter S. Lasecki, Kyle I. Murray, Samuel White, Robert C. Miller, Jeffrey P. Bigham. *Real-time crowd control of existing interfaces*, UIST '11 Proceedings of the 24th annual ACM symposium on User interface software and technology, Pages 23-32 DOI:http://dx.doi.org/10.1145/2047196.2047200

[3] Daren C. Brabham. *Crowdsourcing as a Model for Problem Solving*, The International Journal of Research into New Media Technologies, London, Los Angeles, New Delhi and Singapore Vol 14(1): 75-90 DOI:http://dx.doi.org/10.1177/1354856507084420

[4] Dale Ganleya, Cliff Lampe. *The ties that bind: Social network principles in online communities*, Elsevier, Decision Support Systems, Vol. 47, Issue 3, June 2009, Pages 266-274 DOI:http://dx.doi.org/10.1016/j.dss.2009.02.013

[5] Steven Dow, Anand Kulkarni, Scott Klemmer, Björn Hartmann. *Shepherding the crowd yields better work*, CSCW '12 Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, Pages 1013-1022 DOI:http://dx.doi.org/10.1145/2145204.2145355

[6] Anhai Doan, Raghu Ramakrishnan, Alon Y. Halevy. *Crowdsourcing Systems on the World-Wide Web*, Communications of the ACM Vol. 54 No. 4, Pages 86-96, DOI:http://dx.doi.org/10.1145/1924421.1924442

[7] Kyoung Hoon Yim, MD, Francis Sahngun Nahm, MD, Kyoung Ah Han, Soo Young Park. *Analysis of Statistical Methods and Errors in the Articles Published in the Korean Journal of Pain*, Korean J Pain Vol. 23, No. 1, 2010 DOI:http://dx.doi.org/10.3344/kjp.2010.23.1.35

[8] Vaux, David L. *Research methods: Know when your numbers are significant*, Nature Publishing Group, Vol. 492 No. 7428, Pages 180-181, DOI:http://dx.doi.org/10.1038/492180a

[9] Javier A. Bargas-Avila, Kasper Hornbaek. *Old Wine in New Bottles or Novel Challenges? A Critical Analysis of Empirical Studies of User Experience.* CHI 2011 - Session: user experience, Vancouver, BC, Canada DOI:http://dx.doi.org/10.1145/1978942.1979336

[10] Patrick M. de Boer, Abraham Bernstein, 2014. *PPLib: Towards the Automated Generation of Crowd Computing Programs using Process Recombination and Auto-Experimentation*, ACM Trans. Intelligent Systems and Technology, to be released.

[11] Greg Little, Lydia B. Chilton, Max Goldman, and Robert C. Miller. 2010. *TurKit: human computation algorithms on mechanical turk*, In Proceedings of the 23nd annual ACM symposium on User interface software and technology (UIST '10). ACM, New York, NY, USA, 57-66. DOI:http://doi.acm.org/10.1145/1866029.1866040

[12] Thomas W. Malone et al. 1999. *Toward a handbook of organizational processes*, Management Science 45(3) pp 425-443, March, 1999 DOI:http://dx.doi.org/10.1287/mnsc.45.3.425

[13] Jürgen Angele, Michael Kifer, and Georg Lausen. 2009. Ontologies in F-Logic. In*Handbook on Ontologies*, 45-70. DOI:http://dx.doi.org/10.1007/978-3-540-92673-3_2

[14] Goschin, Sergiu. *Stochastic dilemmas*, Rutgers University-Graduate School-New Brunswick. DOI:http://dx.doi.org/doi:10.7282/T3H993GS

[15] Hospice Houngbo, Robert E. Mercer. *Method mention extraction from scientific research papers*, Department of Computer Science, The University of Western Ontario, London, ON, Canada http://www.aclweb.org/anthology/C12-1074

[16] James Tebby, Fiona Lecky, Antoinette Edwards, Tom Jenks, Omar Bouamra, Rozalia Dimitriou and Peter V. Giannoudis. *Outcomes of polytrauma patients with diabetes mellitus*, Tebby et al. BMC Medicine 2014, http://www.biomedcentral.com/content/pdf/1741-7015-12-111.pdf

[17] Casey L Peiris, Nora Shields, Natasha K Brusco, Jennifer J Watts and Nicholas F Taylor. *Additional Saturday rehabilitation improves functional independence and quality of life and reduces length of stay: a randomized controlled trial*, Peiris et al. BMC Medicine 2013, http://www.biomedcentral.com/content/pdf/1741-7015-11-198.pdf