A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF COMPUTING

DEPARTMENT OF COMPUTER SCIENCE

eAnalysis: Easier statistical analysis

Emile Brès A0132365L

Monday 10th March, 2016

Supervisor : Prof. Wong Limsoon

Acknowledgements

I would first like to express my gratitude to my supervisor, Professor Wong Limsoom, for its continuous support and guidance throughout this project. He was always available to guide me, discuss the issues and correct my mistakes. This project would not have been possible without him. I would also like to thank my family, in particular my father and my girlfriend Camille, that were there to help me when things were hard during this long project. All my love goes to them.

Contents

List of Figures

List of Tables

Introduction

Context We produce more data today than we ever have in human history. According to a study by SINTEF in 2013, 90% of data in world history was created in the last two years [1]. The development of computers and Internet in particular has led to an explosion of the quantity of data produced. Every minute, we are liking 4.2 million posts on Facebook, uploading 300 hours of video to YouTube, and sending 350,000 tweets [2]. The enormous value lying in all this data is well recognized, propelling companies such as Google or Facebook to market caps of hundreds of billions of dollars [3]. Statistical analysis is the main tool to extract information from this huge mass of data and, as such, is becoming more and more important and widely used. Traditionally, statistical analysis was mostly used in a scientific context. Data was collected for a specific purpose (for an experiment for example) in a planned and controlled way. Statistical studies were conducted by statisticians themselves or researchers with training in statistics. However, with the raise of big data, the situation is completely different. The data is collected systematically, often without a determined purpose. The analyst does not know before the analysis what he is looking for. He is rather looking to discover interesting patterns in the data and use them to infer useful properties. Furthermore, scientists are no longer the only ones who have to conduct such analyses. Data analysis has become essential to a large number of companies and many of the people expected to conduct statistical analysis have little training in statistics.

Problems This situation can lead to a number of problems and in particular to a misuse of statistical tests. In scientific experiments, scientists take special care of the way they collect their data. They choose their subjects randomly and study them in a controlled setting to be assured of obtaining data of good quality. This selection is essential to the proper conduct of statistical tests. In fact, statistical tests need to verify several hypotheses in order to be valid. For example, the samples must be independent and identically distributed (i.i.d). However the data usually encountered in a big data context has not been produced in such a controlled environment. There is no warranty that the necessary hypotheses will be verified which can lead to false results. Another problem is that statistical theory is often disregarded in the big data world. It can be perceived as too complicated or irrelevant to the big data workflow where the focus is primarily on the discovery of insights. For example, Mike Flowers, a lawyer, and NYC's first Director of Analytics said in a interview: "I had no interest in very experienced statisticians. I was a little concerned that they would be reluctant to take this novel approach to problem solving. Earlier, when I had interviewed traditional stats guys for the financial fraud project, they had tended to raise arcane concerns about mathematical methods. I wasn't even thinking about what model I was going to use. I wanted actionable insight, and that was all I cared about" [4]. This disregard for statistical theory is problematic because the consequences of failing to conduct a statistical analysis properly can be severe. It is indeed very easy to infer erroneous conclusions from statistics as demonstrated in the famous book *How to Lie with Statistics* [5], which relates a lot of common errors and misconceptions about statistics such as sampling bias or the equivalence of correlation and causality. Statistical results can also be very misleading if they are not interpreted correctly. For example, a relationship between two variables can be present in each different group and be reversed if you combine all the groups. This is known as the Simpson paradox [6]. A good example is the Berkeley gender bias case [7], where the data at the university level showed a bias against women in the graduate school admission, but, for each department, the bias was reversed in favor of women. This paradox occurs because of different distributions of samples in each group and illustrates the importance of properly assessing data hypotheses when conducting an analysis.

Needs However, not everyone has the time or the will to undertake a training in statistics. Sometimes the data is confidential or there is no time to consult an expert. And even experimented users can make mistakes if they are facing a lot of pressure. We thus think that there is a need for a tool that could help a user to conduct and report a statistical analysis properly. Such a tool would need several properties. A first basic functionality would be to be able to run statistical tests and report the results. But it should also be self diagnosing, meaning that it should be able to verify tests hypotheses and select the most appropriate test available in the system automatically. It should also offer additional capabilities such as way for the user to investigate the data he wants to study, to help him construct interesting hypotheses before testing them. Furthermore we would like a tool that not only test an hypothesis but also help the user to infer more insights by conducting additional analysis. This additional analysis would be conducted automatically and reported to the user clearly so that he can use it in its study. Moreover, such a tool should be pedagogical. Since it is primarily targeted to non expert statisticians, it is important that the tool exposes clearly and explains the statistical process it is conducting and the reasons for the choices that it makes during the study. This is also very important to allow the user to communicate the results of the tools to other statisticians through a paper or discussion.

1 Workflow in statistical analysis

Conducting a statistical analysis is not a simple task. A researcher must not only find the most appropriate test to apply to the data and obtain an interesting result. He must also be able to communicate the data and the statistical method he followed so that other researchers can reproduce its results. In fact, a study whose results cannot be reproduced does not have much value and its authors can be considered as incompetent or even dishonest, as the controversy on the water memory has shown [8]. To obtain good results and achieve reproducibility of the results, a good workflow is essential. Researchers have written books explaining how to construct a good workflow, for example Kirchkamp's *Workflow of statistical data analysis* [9] or Long's *The Workflow of Data Analysis Using Stata* [10]. We think that, by automating some parts of the workflow of a statistical study, eAnalysis should help researchers to obtain correct and reproducible results.

In this section, we will describe how a statistical analysis is conducted from the collection of the data to the publication of the results. We will first describe the traditional way in a research setting. We will then see how the data analysis workflow is different in a big data setting.

1.1 Data analysis in a research setting

In a research setting, the statistical analysis is only a part of the study. Typically the researcher has a question in its domain (biology or social sciences), he designs an experiment to answer this question, collects the data and then conducts the statistical analysis. In the Handbook of Biological Statistics [11], McDonald describes a systematic approach to the analysis of biological data:

1. Specify the biological question you are asking.

2. Put the question in the form of a biological null hypothesis and alternate hypothesis.

3. Put the question in the form of a statistical null hypothesis and alternate hypothesis.

4. Determine which variables are relevant to the question.

5. Determine what kind of variable each one is.

6. Design an experiment that controls or randomizes the confounding variables.

7. Based on the number of variables, the kinds of variables, the expected fit to the parametric assumptions, and the hypothesis to be tested, choose the best statistical test to use.

8. If possible, do a power analysis to determine a good sample size for the experiment.

9. Do the experiment.

10. Examine the data to see if it meets the assumptions of the statistical test you chose (primarily normality and homoscedasticity for tests of measurement variables). If it doesn't, choose a more appropriate test.

11. Apply the statistical test you chose, and interpret the results.

12. Communicate your results effectively, usually with a graph or table.

McDonald focuses on biology in his book but this method is valid for every scientific domain. It is important to note that, in a research setting, the statistical analysis itself is just a part of a bigger process. In particular, a lot of work is necessary in order to design and realize an experiment. The data collected must fit a set of conditions in order for the test to be valid and it is the responsibility of the researcher to make sure that these conditions are verified.

1.2 Data analysis in a big data setting

Data analysis in a big data context presents different challenges. In fact, new methods have been designed specifically for big data analysis [12]. Counter to a research setting, the user does not necessarily have a question to ask a priori. The data was often not collected in a controlled manner and with a specific goal and the user can make no assumptions on the shape or distribution of the data. The goal of the data scientist is rather to try to detect interesting patterns and use them to infer valuable insights. A good example of this approach is a project from John Hopkins University that aims to better identify the causes of youth obesity in Pennsylvania by collecting various types of information and detecting unexpected correlations [13]. They do not know in advance the variables who are going to be correlated with obesity but rather are looking for them. In this situation, one of the first steps in analyzing a dataset is to get a general idea of its properties. A popular approach to do so is called exploratory data analysis (EDA) [14] and relies on graphical methods as opposed to statistical hypothesis testing. There are also technical challenges associated with the collection, the curation and the storage of data but they are not in the scope of our study. A data scientist has thus different needs than a scientific researcher. He needs to be able to explore the data easily, by constructing simple measures, such as means, medians and variances and by drawing graphs. He also needs a system powerful enough to handle datasets with large numbers of values.
The workflow of a data analyst could be the following:

1. Load the data

2. Compute some general metrics (mean, median, variance) and plot a few graphs to get a general idea of the properties of the dataset

3. Generate statistical models of the data and conduct statistical tests

4. Iterate

5. Communicate the results

Data scientists are generally less concerned with carefully assessing that specific statistical hypotheses are verified and more focused on detecting patterns and obtaining interesting insights.

2 eAnalysis

2.1 System presentation

To address the needs described in the introduction and building on the reflection described in Section 1, we propose a tool, named eAnalysis, which is a contraction of **ea**sy **a**nalysis. eAnalysis is a web application that aims to make statistical analysis easier by automating important parts of the statistical process. eAnalysis covers the complete statistical process from the loading of data to the computation of the test and the publication of the results. eAnalysis allows the user to load a dataset, explore it and select the variables he is interested in. The system will then assess if the selected variables are associated with each other. To do so, eAnalysis automatically selects and computes the most appropriate statistical test available in the system. Furthermore eAnalysis conducts additional analysis automatically to help the user gain more insights.

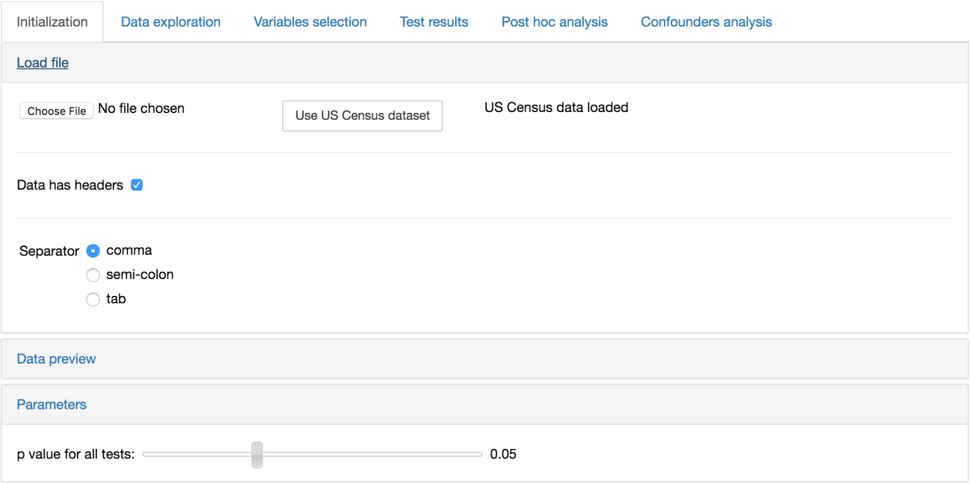
Statistical analysis is a complex subject. Thus, one of eAnalysis main concerns was to be as simple and accessible as possible. In order to do so, eAnalysis is distributed primarily as a web application. This allows users to use eAnalysis without needing to install any statistical package. An online version of eAnalysis can be found at `128.199.231.116:8000/notebooks/easy_analysis.ipynb`. A special care was also taken to make the graphical interface as intuitive and natural as possible. Finally, eAnalysis provides detailed explanations in plain English at each step of the process so that the user can understand how the system works.

In this section, we will present in details the interface of eAnalysis, the statistical tools that it offers, the technologies used to build the system and the performances obtained.

2.2 User interface

2.2.1 Walkthrough

eAnalysis is organized into several tabs, with each tab housing a specific functionality. This separation allows a clearer workflow for the user by representing clearly each step of the statistical process. The user interface acts thus as a canvas and a support for the statistical process.



No file chosen

US Census data loaded

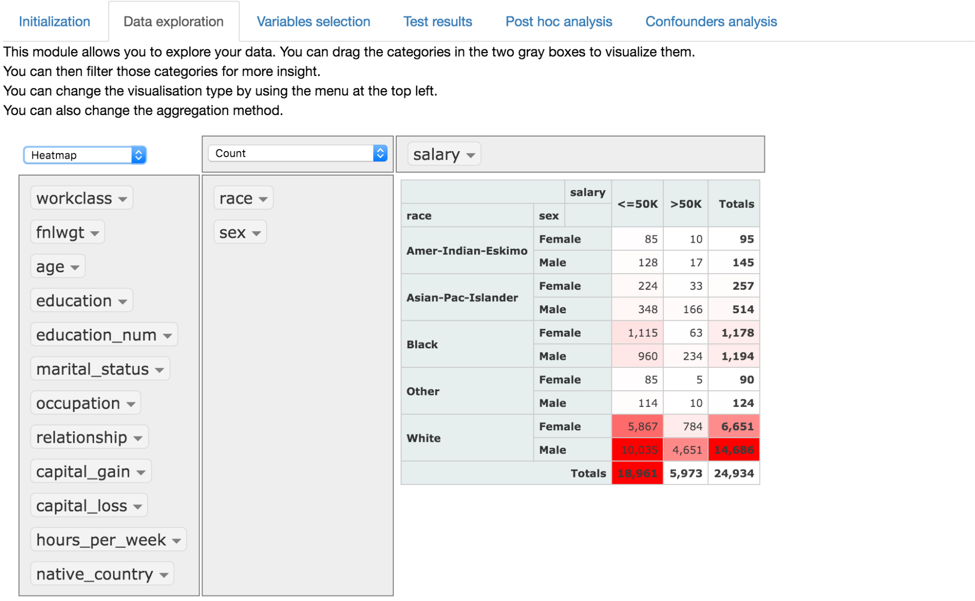
Data has headers ☑

Separator ● comma
○ semi-colon
○ tab

p value for all tests: 0.05

Figure 1: Initialization of the study: data file loading and parameter settings

Initialization The first tab *Initialization* (Figure 1) corresponds to the initialization of the study. The user can load the data into the system via a dedicated button by selecting a file on his computer. He can also choose to use the adults dataset of the UCI Machine Learning Repository as an example dataset [15]. The user can then specify some parameters in order to load the data file correctly: he can specify if the file has a headers row and choose the appropriate separator between comma (a classic CSV file), semicolon or tab (a TSV file). The user can then see a preview of the data. It allows him to assess if the data is correctly loaded in the system. Finally the user can also specify the parameters of the statistical study, namely the p value used throughout the study. The p value chosen here will be used for every statistical test.

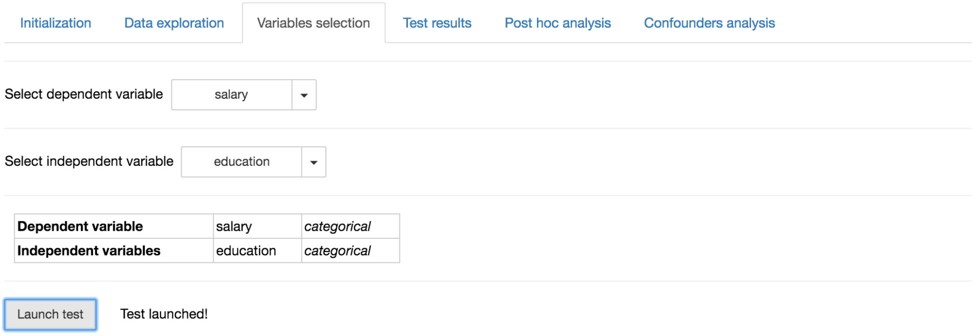


This module allows you to explore your data. You can drag the categories in the two gray boxes to visualize them.
You can then filter those categories for more insight.
You can change the visualisation type by using the menu at the top left.
You can also change the aggregation method.

		salary	<=50K	>50K	Totals
race	sex				
Amer-Indian-Eskimo	Female		85	10	95
	Male		128	17	145
Asian-Pac-Islander	Female		224	33	257
	Male		348	166	514
Black	Female		1,115	63	1,178
	Male		960	234	1,194
Other	Female		85	5	90
	Male		114	10	124
White	Female		5,867	784	6,651
	Male		10,035	4,651	14,686
		Totals	18,961	5,973	24,934

Figure 2: Exploration of the data: Heatmap representing the distributions of salary groups according to race and sex

Data exploration The second tab *Data exploration* (Figure 2) allows the user to explore the data interactively via a pivot table. The user can select the variables that interest him and present them in every possible combination. Several aggregation methods are available: count, average, maximum, minimum, etc. For example, for the adults dataset, the user can see the distributions of subjects earning more or less than 50K per year according to their race and sex. He can also filter the data on each variable to gain further insights. Moreover, several visualizations are available from table to heatmap and bar charts. The versatility of this tool offers many possibilities for the user to obtain a first knowledge of the data. This will help him detect interesting patterns and guide its decision into what statistical tests to conduct.



Select dependent variable: salary

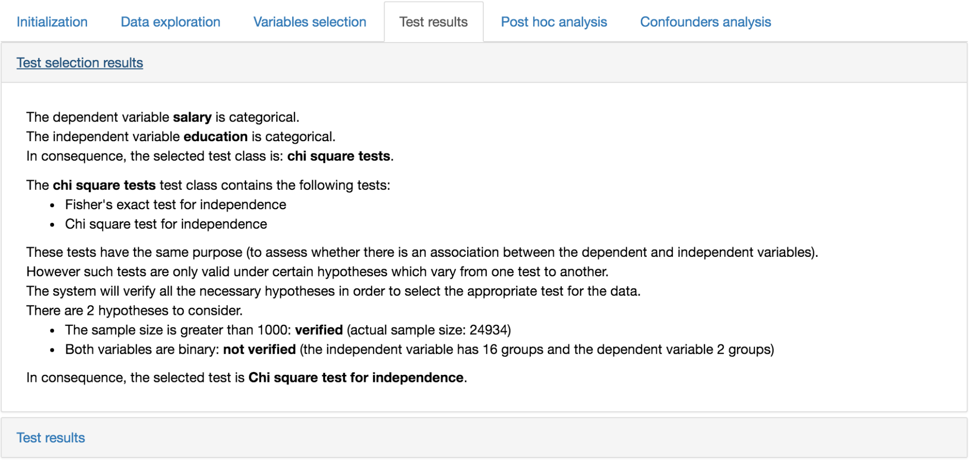
Select independent variable: education

Dependent variable	salary	*categorical*
Independent variables	education	*categorical*

Test launched!

Figure 3: Selection of the test variables and display of the variable types

Variables selection The third tab *Variables selection* (Figure 3) allows the user to select the variables that will be studied by the system. The user chooses one dependent and one independent variable. The dependent variable is the outcome of the study, the variable whose variation is being studied. The independent variable is the input of the study or the potential cause for variations of the dependent variable. When the user selects the variables, the system displays their type. They can be categorical (i.e. the values of the variable are part of a finite set) or numerical (i.e. the values of the variables are numerical). This distinction is important because it determines the type of statistical tests that will be applied. The user can then launch the test by clicking on the *Launch test* button. If the test starts correctly, a confirmation message is displayed next to the button.



The dependent variable **salary** is categorical.
The independent variable **education** is categorical.
In consequence, the selected test class is: **chi square tests**.

The **chi square tests** test class contains the following tests:

- Fisher's exact test for independence
- Chi square test for independence

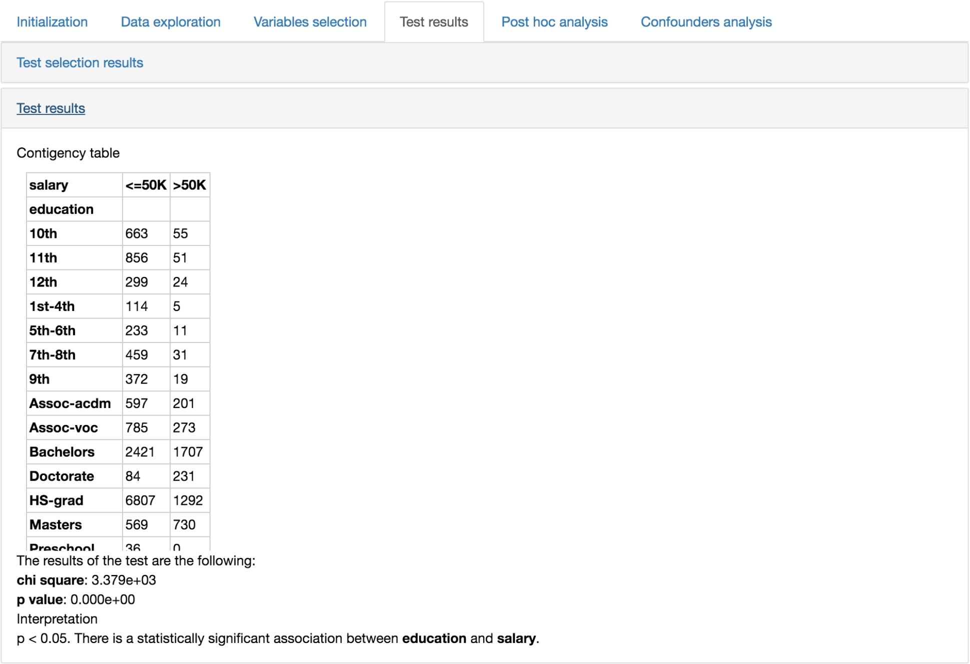
These tests have the same purpose (to assess whether there is an association between the dependent and independent variables).
However such tests are only valid under certain hypotheses which vary from one test to another.
The system will verify all the necessary hypotheses in order to select the appropriate test for the data.
There are 2 hypotheses to consider.

- The sample size is greater than 1000: **verified** (actual sample size: 24934)
- Both variables are binary: **not verified** (the independent variable has 16 groups and the dependent variable 2 groups)

In consequence, the selected test is **Chi square test for independence**.

Figure 4: Detailed explanations of the test selection

Test results The fourth tab *Test results* has two parts. The first part *Test selection results* (Figure 4) displays the information regarding the selection of the appropriate statistical test. This selection is done in two steps: the selection of the test class and the selection of the test itself. eAnalysis calls test class a set of statistical tests which apply to the same type of variables (categorical or numerical) and consider the same null hypothesis. However the tests in a test class differ by the hypotheses that are necessary to conduct the analysis. Some tests require strong hypotheses (the t test requires the normality of data distribution for example) while others have less strict hypotheses. A test whose hypotheses are not verified will give false results. We could think that a way to choose a test is thus to always use the test with the weakest hypotheses. However tests with weaker hypotheses have often smaller statistical power, which can lead to Type 1 errors and weaken the analysis. Therefore it is important to select the appropriate test for each situation. eAnalysis first selects the test class in function of the type of the dependent and independent variables. Then it checks all the hypotheses necessary in order to select the most appropriate test available in the system. The results of the hypothesis checks are displayed to allow the user to understand the reasons why a particular test was chosen. The test selection process and the available tests are presented in more details in Section 2.3.



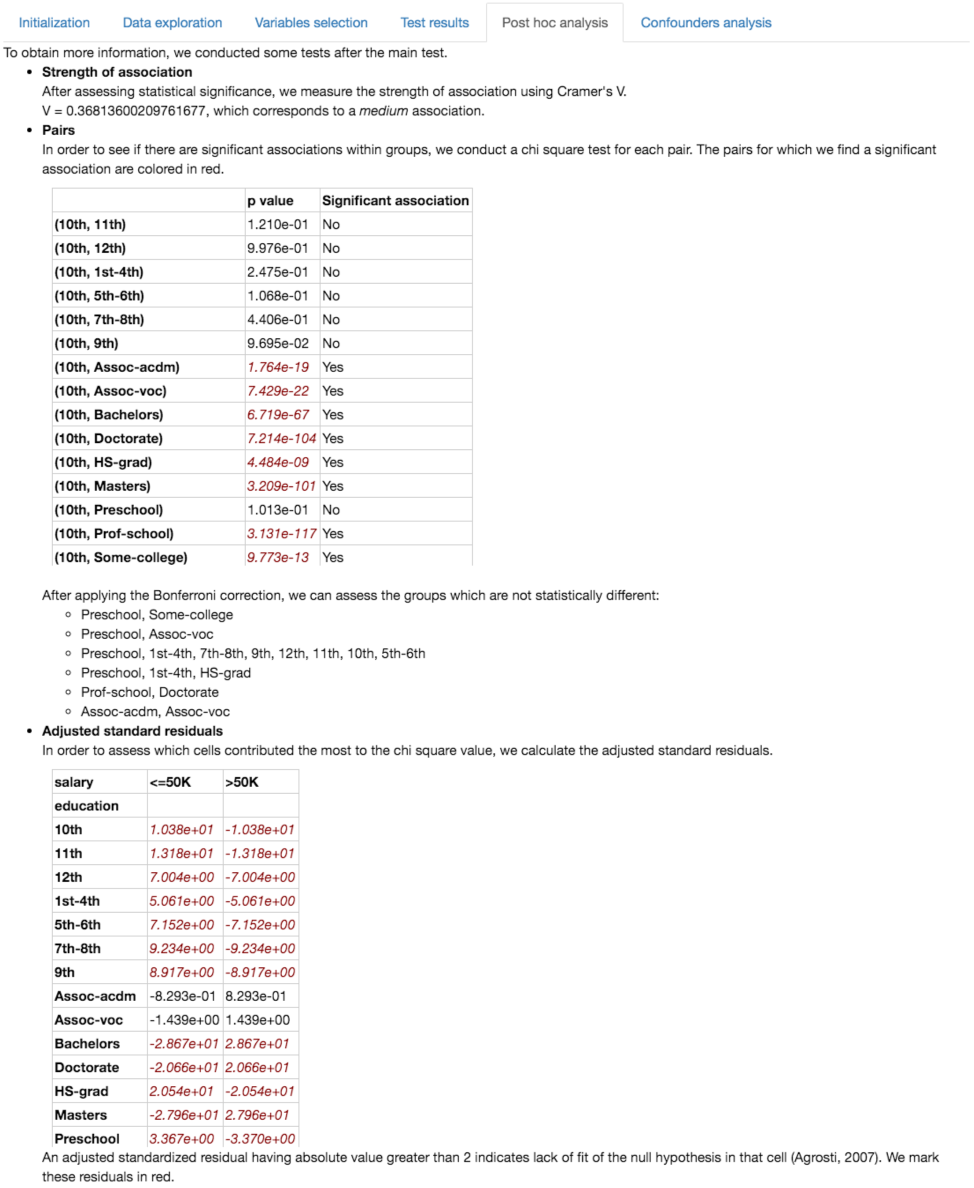
Contigency table

salary	<=50K	>50K
education		
10th	663	55
11th	856	51
12th	299	24
1st-4th	114	5
5th-6th	233	11
7th-8th	459	31
9th	372	19
Assoc-acdm	597	201
Assoc-voc	785	273
Bachelors	2421	1707
Doctorate	84	231
HS-grad	6807	1292
Masters	569	730
Preschool	36	0

The results of the test are the following:
chi square: 3.379e+03
p value: 0.000e+00
Interpretation
p < 0.05. There is a statistically significant association between **education** and **salary**.

Figure 5: Results of the tests with simple interpretation and contingency table

The second part *Test results* (Figure 5) presents the actual results of the test selected by eAnalysis. It displays the value of the test, the p value and an interpretation of the result. It will also display some additional information about the test, such as a contingency table, if relevant. This part of the tab has all the information that a researcher needs to write in a study when presenting a statistical result.



To obtain more information, we conducted some tests after the main test.

- **Strength of association**
  After assessing statistical significance, we measure the strength of association using Cramer's V.
  V = 0.36813600209761677, which corresponds to a *medium* association.
- **Pairs**
  In order to see if there are significant associations within groups, we conduct a chi square test for each pair. The pairs for which we find a significant association are colored in red.

	p value	Significant association
(10th, 11th)	1.210e-01	No
(10th, 12th)	9.976e-01	No
(10th, 1st-4th)	2.475e-01	No
(10th, 5th-6th)	1.068e-01	No
(10th, 7th-8th)	4.406e-01	No
(10th, 9th)	9.695e-02	No
(10th, Assoc-acdm)	*1.764e-19*	Yes
(10th, Assoc-voc)	*7.429e-22*	Yes
(10th, Bachelors)	*6.719e-67*	Yes
(10th, Doctorate)	*7.214e-104*	Yes
(10th, HS-grad)	*4.484e-09*	Yes
(10th, Masters)	*3.209e-101*	Yes
(10th, Preschool)	1.013e-01	No
(10th, Prof-school)	*3.131e-117*	Yes
(10th, Some-college)	*9.773e-13*	Yes

After applying the Bonferroni correction, we can assess the groups which are not statistically different:

  - Preschool, Some-college
  - Preschool, Assoc-voc
  - Preschool, 1st-4th, 7th-8th, 9th, 12th, 11th, 10th, 5th-6th
  - Preschool, 1st-4th, HS-grad
  - Prof-school, Doctorate
  - Assoc-acdm, Assoc-voc

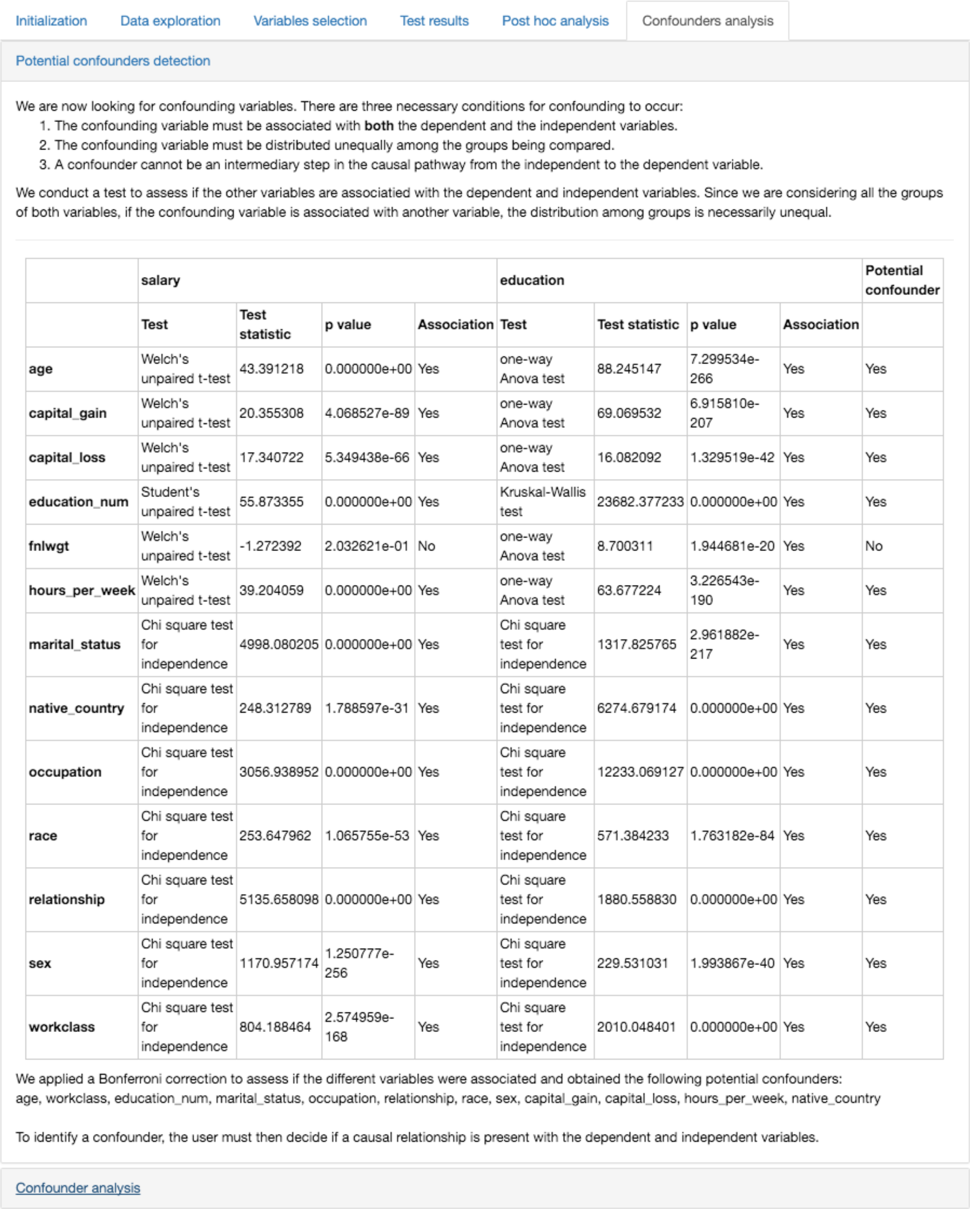
- **Adjusted standard residuals**
  In order to assess which cells contributed the most to the chi square value, we calculate the adjusted standard residuals.

salary	**<=50K**	**>50K**
education		
10th	*1.038e+01*	*-1.038e+01*
11th	*1.318e+01*	*-1.318e+01*
12th	*7.004e+00*	*-7.004e+00*
1st-4th	*5.061e+00*	*-5.061e+00*
5th-6th	*7.152e+00*	*-7.152e+00*
7th-8th	*9.234e+00*	*-9.234e+00*
9th	*8.917e+00*	*-8.917e+00*
Assoc-acdm	-8.293e-01	8.293e-01
Assoc-voc	-1.439e+00	1.439e+00
Bachelors	*-2.867e+01*	*2.867e+01*
Doctorate	*-2.066e+01*	*2.066e+01*
HS-grad	*2.054e+01*	*-2.054e+01*
Masters	*-2.796e+01*	*2.796e+01*
Preschool	*3.367e+00*	*-3.370e+00*

An adjusted standardized residual having absolute value greater than 2 indicates lack of fit of the null hypothesis in that cell (Agrosti, 2007). We mark these residuals in red.

Figure 6: Post hoc analysis: calculation of the strength of association, assessment of the association between groups of the independent variable and calculation of the adjusted standard residuals

Post hoc analysis eAnalysis also provides additional information by running more tests, called post hoc tests. The post hoc tests are selected automatically based on the types of the test variables and their results are presented in the fifth tab *Post hoc analysis* (Figure 6). The post hoc tests are presented in more details in Section 2.3.



We are now looking for confounding variables. There are three necessary conditions for confounding to occur:

1. The confounding variable must be associated with **both** the dependent and the independent variables.
2. The confounding variable must be distributed unequally among the groups being compared.
3. A confounder cannot be an intermediary step in the causal pathway from the independent to the dependent variable.

We conduct a test to assess if the other variables are associatied with the dependent and independent variables. Since we are considering all the groups of both variables, if the confounding variable is associated with another variable, the distribution among groups is necessarily unequal.

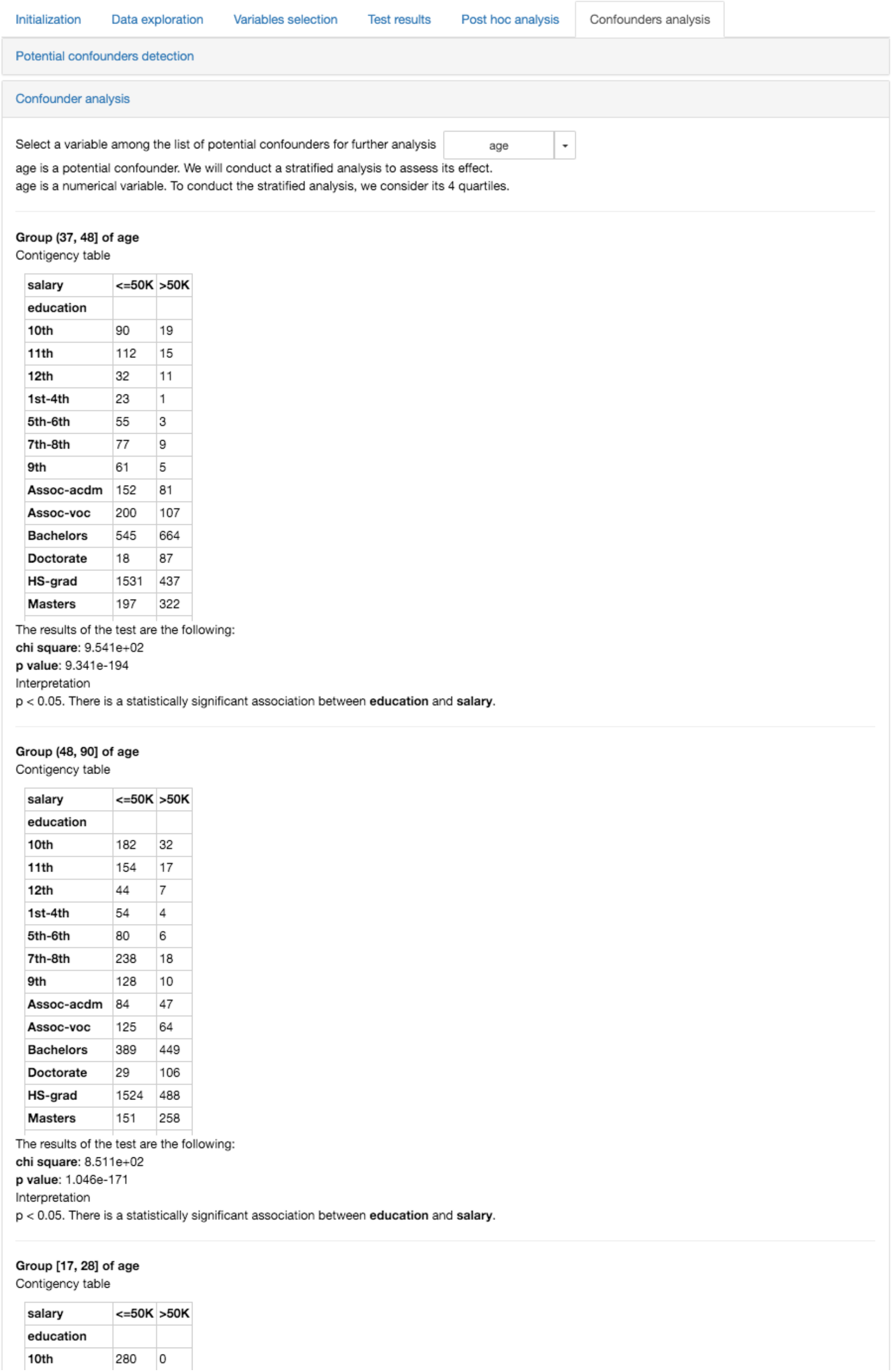
	salary				**education**				**Potential confounder**
	Test	**Test statistic**	**p value**	**Association**	**Test**	**Test statistic**	**p value**	**Association**	
age	Welch's unpaired t-test	43.391218	0.000000e+00	Yes	one-way Anova test	88.245147	7.299534e-266	Yes	Yes
capital_gain	Welch's unpaired t-test	20.355308	4.068527e-89	Yes	one-way Anova test	69.069532	6.915810e-207	Yes	Yes
capital_loss	Welch's unpaired t-test	17.340722	5.349438e-66	Yes	one-way Anova test	16.082092	1.329519e-42	Yes	Yes
education_num	Student's unpaired t-test	55.873355	0.000000e+00	Yes	Kruskal-Wallis test	23682.377233	0.000000e+00	Yes	Yes
fnlwgt	Welch's unpaired t-test	-1.272392	2.032621e-01	No	one-way Anova test	8.700311	1.944681e-20	Yes	No
hours_per_week	Welch's unpaired t-test	39.204059	0.000000e+00	Yes	one-way Anova test	63.677224	3.226543e-190	Yes	Yes
marital_status	Chi square test for independence	4998.080205	0.000000e+00	Yes	Chi square test for independence	1317.825765	2.961882e-217	Yes	Yes
native_country	Chi square test for independence	248.312789	1.788597e-31	Yes	Chi square test for independence	6274.679174	0.000000e+00	Yes	Yes
occupation	Chi square test for independence	3056.938952	0.000000e+00	Yes	Chi square test for independence	12233.069127	0.000000e+00	Yes	Yes
race	Chi square test for independence	253.647962	1.065755e-53	Yes	Chi square test for independence	571.384233	1.763182e-84	Yes	Yes
relationship	Chi square test for independence	5135.658098	0.000000e+00	Yes	Chi square test for independence	1880.558830	0.000000e+00	Yes	Yes
sex	Chi square test for independence	1170.957174	1.250777e-256	Yes	Chi square test for independence	229.531031	1.993867e-40	Yes	Yes
workclass	Chi square test for independence	804.188464	2.574959e-168	Yes	Chi square test for independence	2010.048401	0.000000e+00	Yes	Yes

We applied a Bonferroni correction to assess if the different variables were associated and obtained the following potential confounders:
age, workclass, education_num, marital_status, occupation, relationship, race, sex, capital_gain, capital_loss, hours_per_week, native_country

To identify a confounder, the user must then decide if a causal relationship is present with the dependent and independent variables.

Figure 7: Confounders analysis: Detection of potential confounders

Confounders analysis The sixth tab *Confounders analysis* allows the user to search for potential confounders in the dataset and to assess their impact on the relationship between the independent and dependent variables. The tab is split in two parts. In the first part (Figure 7), eAnalysis computes the association between the dependent variable and every other variable (except the independent variable) on one hand and the association between the independent variable and every other variable (except the dependent variable) on the other hand. If the associations between a variable and both the dependent and independent variables are significant, then this variable is a potential confounder. The user can then use the second part of the tab to assess the effect of that potential confounder by conducting a stratification analysis. If the potential confounder is a numerical variable, eAnalysis will use its 4 quartiles to conduct the stratification, as shown for the variable *age* in Figure 8.



Select a variable among the list of potential confounders for further analysis age

age is a potential confounder. We will conduct a stratified analysis to assess its effect.
age is a numerical variable. To conduct the stratified analysis, we consider its 4 quartiles.

Group (37, 48] of age
Contigency table

salary	<=50K	>50K
education		
10th	90	19
11th	112	15
12th	32	11
1st-4th	23	1
5th-6th	55	3
7th-8th	77	9
9th	61	5
Assoc-acdm	152	81
Assoc-voc	200	107
Bachelors	545	664
Doctorate	18	87
HS-grad	1531	437
Masters	197	322

The results of the test are the following:
chi square: 9.541e+02
p value: 9.341e-194
Interpretation
p < 0.05. There is a statistically significant association between **education** and **salary**.

Group (48, 90] of age
Contigency table

salary	<=50K	>50K
education		
10th	182	32
11th	154	17
12th	44	7
1st-4th	54	4
5th-6th	80	6
7th-8th	238	18
9th	128	10
Assoc-acdm	84	47
Assoc-voc	125	64
Bachelors	389	449
Doctorate	29	106
HS-grad	1524	488
Masters	151	258

The results of the test are the following:
chi square: 8.511e+02
p value: 1.046e-171
Interpretation
p < 0.05. There is a statistically significant association between **education** and **salary**.

Group [17, 28] of age
Contigency table

salary	<=50K	>50K
education		
10th	280	0

Figure 8: Confounders analysis: Stratification analysis with age

2.2.2 How the workflow helps the user

As described in Section 2.2, the workflow of eAnalysis is as follows:

- The user loads the data
- The user explores the data using a tool offering filter, visualization and aggre-

gation

- The user selects the variables he wants to study
- The system automatically selects and computes the most appropriate statistical test available in the system
- The system conducts additional statistical analysis
  - Analysis of the different groups of each variables
  - Analysis of potential confounders and their effect

The workflow is designed to be coherent with the process of real statisticians. It follows in particular the best practices described in the Handbook of biological statistics [11]. Thanks to this approach, a user with even a small experience in statistical analysis will be familiar with the workflow and will be able to understand easily how eAnalysis works. In case the user needs help from a trained statistician, it will also be easier for the statistician to grasp quickly the rationale of the study. However, automating the statistical analysis process can also create problems. In fact, there is a danger that the user uses blindly the tool without understanding the reasoning behind the choices that the system makes. To address this problem, eAnalysis displays detailed explanations at each step of its process. This way the user can follow the reasoning of the system and, if necessary, can report the results in a paper or a report by copying the explanations offered by the system.

eAnalysis can be used both in a research context or in a big data context. A researcher will particularly benefit from the assurance that the selected test is correct, while a big data analyst will be able to use the powerful and versatile tools offered by eAnalysis to explore the data and gain more insights on his hypotheses thanks to the post hoc analysis. Moreover, even if researchers and data scientists may have different approaches regarding data analysis, the tools offered can be useful in both settings. A researcher might discover new hypotheses to test using the data exploration tool and a data scientist can benefit from the automatic selection of a statistical test.

2.3 Statistical tools

In this section, we will present the statistical tests that are available in eAnalysis along with the statistical assumptions that need to be assessed in order to choose the most appropriate test available in the system. We will also explain how eAnalysis chooses the test to run and the post hoc tests.

2.3.1 Structure of statistical tools

One of the main goal of eAnalysis is, given a dataset and a hypothesis to test, to select automatically the most appropriate test available in the system. In order to select the most appropriate test, there are two main criteria. The first criterion is the number and type (numerical or categorical) of the variables of the dependent and the independent variables. This first criterion gives us a set of available statistical tests who can be used to answer the question. These tests typically have different assumptions and different statistical powers. We call such a set a statistical test class or simply test class. The concept of statistical test class is useful both to help the user understand how eAnalysis selects a test and for development. On the development side, using statistical test classes allowed us to give a modular structure to eAnalysis. It is thus very easy to add new statistical tests in an existing test class or to create a new test class, but also to add new assumptions or change the output. The second criterion is to assess which assumptions are valid and which are not. This will allow us to choose the most appropriate statistical test available in the system. However it should be noted that the list of assumptions in each test class is incomplete. In fact, some assumptions can be difficult or impossible to test by using only the data. For example, to conduct a statistical test, the observations should be independent from one another. However, this cannot be determined solely by studying the data as it depends of the way the experience was designed or the way the data was collected. A statistical test class contains several statistical tests and several assumptions. eAnalysis assesses every assumption of the selected test class and then chooses the most powerful test within the subset of tests whose assumptions are valid. Usually this subset contains only one element as the tests have been designed to be the most powerful possible under certain assumptions. The concept of statistical test classes is, to our knowledge, unique to eAnalysis. In order to build the test classes, we made a synthesis of several resources to determine the appropriate tests and statistical assumptions [11, 16, 54, 17]. Statistical test classes are not the only statistical tool in eAnalysis which is modular. In fact, the tests and the statistical assumptions are themselves modular. The modular structure of eAnalysis allows a developer to easily understand how the system is organized and how to extend it. This is very important for us as eAnalysis is an open source project and is destined to be supported by a community.

2.3.2 Statistical test classes

In this section, we will present the available statistical test classes, detailing the tests, assumptions and post hoc analysis in each class. All the statistical tests take for null hypothesis: *the dependent variable(s) and the independent variable(s) are not associated.*

The table 1 shows how the statistical test classes are associated with the variables of interest. The name of the test class is typically the name of the most common test in the test class.

Dependent variable(s)	**Independent variable(s)**	**Test class**
1 categorical	1 categorical	Chi square tests
1 numerical	1 categorical (binary)	T tests
1 numerical	1 categorical (multi level)	ANOVA tests
1 numerical	1 numerical	Linear regression tests

Table 1: Association between dependent and independent variables and test classes

Chi square tests The chi square test class is used when there are one categorical dependent variable and one categorical independent variable. The test class contains two tests: Fisher exact test [18] and Pearson chi square test [19]. The statistical assumptions to assess are:

- The sample size is less than 1000.
- The contingency table has two columns and two rows.

If both assumptions are verified, Fisher exact test is used. Otherwise it is Pearson chi square test. In fact, even though there exists a generalized version of Fisher exact test for multi level variables [20], there is no implementation of this algorithm in Python. Moreover, Fisher exact test is computationally expensive. Thus, if the sample size is too great, we prefer to use the chi square test.

For the post hoc analysis, the chi square test class contains several tools. First eAnalysis estimates the strength of the association between the variables by calculating Cramer's V [21], using a custom implementation. Then, if the independent variable contains more than two groups, the system analyses the differences in the distribution of the dependent variable for each pair by conducting a chi square or Fisher exact test (in fact eAnalysis uses the assumptions of the chi square test class to determine the most appropriate test) on the data filtered down to the two groups of each pair and then applying a Bonferroni correction [22] to correct for the fact that several tests are being conducted, thus raising the probability of a false positive. Using this information, eAnalysis can build a graph representing each group as a vertex, linked by an edge if the pair has similar distributions in each group. The system then computes the sets of groups whose distributions are similar by using a graph theory algorithm for calculating the maximal cliques in a graph [23, 24].

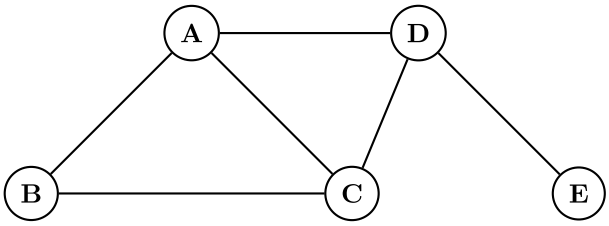


Figure 9: Graph representing the five groups of an independent variable. An edge links the groups for whom the distributions of the dependent variable are similar.

For example, A, B, C, D and E are five groups of the independent variable and the distribution of the dependent variable is independent from the dependent variable for the pairs (A,B), (A,C), (B,C), (A,D), (C,D) and (D,E). We can then draw a graph where two groups share an edge if their distributions are similar, as shown in Figure 9. We then build the maximal sets for which all items in the set have similar distributions. If we start with A, A has an edge to B, C and D. We form a set [A,B] and try to see if we can add more items to the set. C shares an edge with both A and B and can then be added to the set, which gives us [A,B,C]. We try to see if we can add more items. D has an edge with A and C but not with B. Thus D cannot be added to the set. E has no edge to either A, B or C and thus cannot be added to the set. There are no more vertices, thus [A, B, C] is a maximal set of items with similar distributions. By repeating the same process for each vertex, we obtain the sets [A,B,C], [A,C,D] and

[D,E]. Finally, eAnalysis calculates the adjusted standard residuals [25, 26] to assess which group contributed the most to the chi square value.

T tests The T test class is used when there are one numerical dependent variable and one binary independent variable. The test class contains three tests: Student's T test [27], Welch's T test [28] and Kruskal-Wallis test [29]. The statistical assumptions to assess are:

- The distributions of the dependent variable are normal for all groups of the independent variable.
- The variances of the distribution of the dependent variable are homogeneous between all groups of the independent variable (homoscedasticity).
- The distributions of the dependent variable for each group have all their skewnesses of the same sign on one hand and all their kurtosises of the same sign on the other hand. Expressed less formally, it means that they deviate from normal distributions in the same direction.

The Figure 10 shows the decision tree for selecting the most appropriate test available in the system. The decision tree was mostly inspired by McDonald's handbook [11].

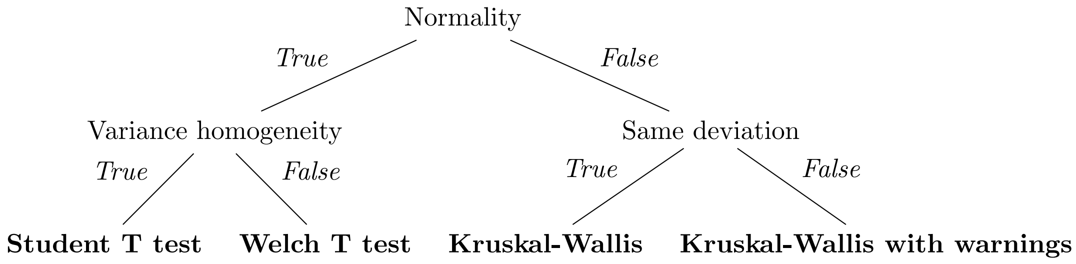


Figure 10: Decision tree for T tests class

In the case, where the distributions are not normal and do not deviate in the same direction, we did not find a perfectly appropriate test in the literature. So we chose to use the Kruskal-Wallis test as the most suitable implemented in eAnalysis but to also display a warning.

ANOVA tests The ANOVA test class is used when there are one numerical dependent variable and one multi level categorical independent variable. The ANOVA test class is a generalization of the T test class presented in Section 2.3.2 and we could have used it to implement the T test class. However, since T tests are commonly used, we decided to keep them separated in order to be more familiar to users. The test class contains two tests: the one-way ANOVA test [30] and Kruskal-Wallis test [29]. The statistical assumptions to assess are:

- The distributions of the dependent variable are normal for all groups of the independent variable.
- The variances of the distribution of the dependent variable are homogeneous between all groups of the independent variable (homoscedasticity).

- The distributions of the dependent variable for each group have all their skewnesses of the same sign on one hand and all their kurtosises of the same sign on the other hand. Expressed less formally, it means that they deviate from normal distributions in the same direction.

The figure 10 shows the decision tree for selecting the appropriate test. The design of the decision tree was mostly inspired by McDonald's handbook [11].

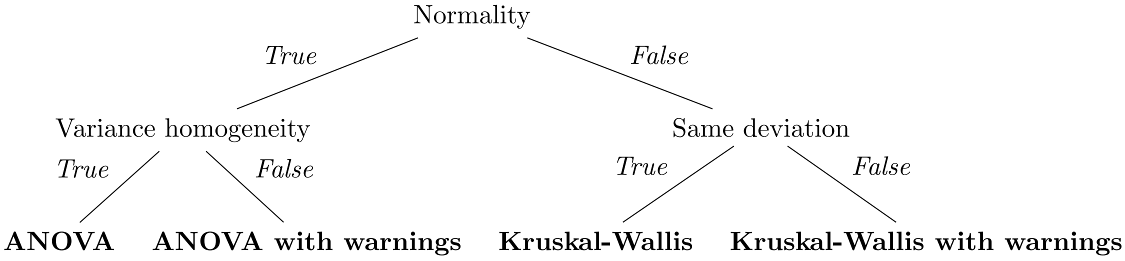


Figure 11: Decision tree for ANOVA test class

Here we have two cases where perfectly appropriate tests could not be found. We proceed in the same way as in Section 2.3.2.

The post hoc analysis follows the same pattern as the chi square test class. If the independent variable contains more than two groups, the system analyses the differences in the distribution of the dependent variable for each pair by conducting a one-way ANOVA test and applying a Bonferroni correction [22]. eAnalysis then computes the maximal sets of groups whose distributions are similar by using a graph theory algorithm for calculating the maximal cliques in a graph [23, 24].

Linear regression tests The linear regression test class is used when there are one numerical dependent variable and one numerical independent variable. The test computes a linear least-squares regression and tests the null hypothesis that the slope is zero. We do not test any assumptions here because there are no alternatives to least-squares regression analysis implemented in scipy. But this is not specific to scipy as Stromberg has shown that the lack of implementation of robust alternatives to the least-squares method is pretty generic [31]. Pending later extension, the linear regression class should be consider as experimental.

2.4 Technologies

Python does not possess an integrated framework for developing web applications, such as Shiny for R [32]. Thus, in order to develop eAnalysis in Python, it was necessary to use and integrate several technologies. In this section, we will present the technological stack we used. It is only composed of free and open source software. As some of the technologies used are still in active development (Jupyter, Docker), the integration was somehow difficult and the stack may be quite hard to use at the moment. But, we think that, as the tools get more stable, such a stack could be integrated to offer a functional equivalent to Shiny for Python. The source code for eAnalysis itself is available at `https://bitbucket.org/emilebres/eanalysis/src`.

An online version of eAnalysis is hosted on a server based in Singapore and owned by Digital Ocean with 1 GB of RAM and can be found at `128.199.231.116:8000/notebooks/easy_analysis.ipynb`.

2.4.1 Python

Python is a widely used high-level, general-purpose, interpreted, dynamic programming language [33]. It is, with R, one of the most popular language for statistics and data analysis. eAnalysis is entirely written in Python [34], the back-end as well as the front-end. Numerous libraries have been developed in Python for a variety of purposes. eAnalysis leverages some of these libraries to perform statistical tasks and display the user interface. pandas is a Python library that provide data structures for data analysis and modelling, designed for ease of use and high performance [35]. eAnalysis uses pandas to store the data in pandas datatables and then manipulate the data. scipy is a Python library for mathematics, science and engineering [36]. All the statistical tests in eAnalysis are implemented using scipy functions as a base.

2.4.2 Jupyter

Jupyter is a web application used to share and display code interactively [37]. It is widely used in the Python data science community and allows users to display source code and analysis results in the same place. Jupyter functionalities can be extended by using plugins and many have been developed by the community. eAnalysis leverages some of them, mostly the Jupyter Notebook Extensions [38] and ipywidgets [39] to transform a Jupyter notebook in a real web application. Such an extensive use of the plugins to develop a full-fledged web application is a fairly new concept for Jupyter and the plugins used are still under heavy development. Thus some effort was necessary in order to integrate the plugins correctly. But, as the tools evolve, it will become easier.

2.4.3 Docker

Docker is a software used to automate the deployment of applications [40]. Docker packages a software and all its dependencies into a light-weight container that can run on any computer. This was very helpful for eAnalysis as the application depends on several software systems. In particular, there are two main versions of Python, several versions of scipy, of pandas and of Jupyter. These versions are not all compatible with each other. By using Docker, we stabilize the software versions and make sure that eAnalysis can run on any platform without trouble. Finally we use a project called jupyter-tmpnb [41] in order to manage and deploy eAnalysis. jupyter-tmpnb launches a container for each user of eAnalysis. This way, all users can run eAnalysis without interfering with each other.

2.4.4 Other technologies

Other technologies are also used in eAnalysis. The pivot table tool used for data exploration comes from a JavaScript library, PivotTable.js [42]. It has been adapted in order to be used in Jupyter and can take pandas datatables as input. We also used HTML and CSS to organize and style the interface.

2.5 Performances

To assess the computational performances of eAnalysis, we conducted tests on three different datasets of the UCI Machine Learning Repository. Their characteristics are presented in table 2.

Dataset	Number of instances	Number of attributes	Attribute types
abalone	4177	8	Categorical, Integer, Real
adult	48842	14	Categorical, Integer
mushroom	8124	22	Categorical

Table 2: Datasets used to assess the performance of eAnalysis

We conducted the tests on two machines: on one side a Digital Ocean droplet, a virtual private server with a processing power equivalent to 1.76 GHz (as measured by Cloudlook [43]) and 1 GB of RAM, on the other side a personal computer with a 2.5 GHZ processor and 8 GB of RAM. For each dataset, we measure the time spent on a complete analysis (test selection and calculation, post hoc analysis). The results are displayed in Table 3. The total duration of the statistical process is never greater than 3 seconds. The performances obtained are thus acceptable for a data analysis tool.

Dataset	Dependent variable	Independent variable	Time on computer (seconds)	Time on server (seconds)
abalone	diameter	sex	0.123	0.224
adult	salary	sex	1.65	2.40
adult	salary	race	0.534	0.840
adult	education	race	0.755	1.70
adult	race	sex	0.469	1.01
mushroom	odor	habitat	0.560	1.14

Table 3: Comparison of duration of the statistical process of eAnalysis across multiple datasets, variables and machines

3 Use case

In this section, we present a typical use case for eAnalysis. To do that, we will use the adults dataset (from the UCI machine learning repository) [15].This use case will serve as an example of eAnalysis functionalities as well as a demonstration of the performances obtained with an analysis for a dataset of 15 columns and approximately 25000 rows. The user wants to study the association between race and salary in the dataset. The parts below will describe the workflow of eAnalysis to address this question.

3.1 Setting up the study

The first step consists of loading the dataset into eAnalysis. If the user has downloaded the dataset beforehand, he loads the file by selecting it in its computer using the *Choose file* button. After checking in the *Data preview* section that the dataset is correctly loaded, the user can change the p value used for each test in the *Parameters* section, the default value being 0.05.

3.2 Exploring the data

Once the data is loaded in the system, the user can use the tool offered in the *Data exploration* tab to explore the data. In this case, the user can represent the count of each category in a table to get a first insight as shown in Table 4.

	<=50K	**>50K**
Amer-Indian-Eskimo	213	27
Asian-Pac-Islander	572	199
Black	2075	297
Other	199	15
White	15902	5435

Table 4: Contingency table of race and salary in the adult dataset

Here he can see that there are much more White than all the other races and that there are much more people earning less than 50K than people earning more. To get a better idea of the distribution, the user can also display the data in a bar chart as shown in the Figure 12.

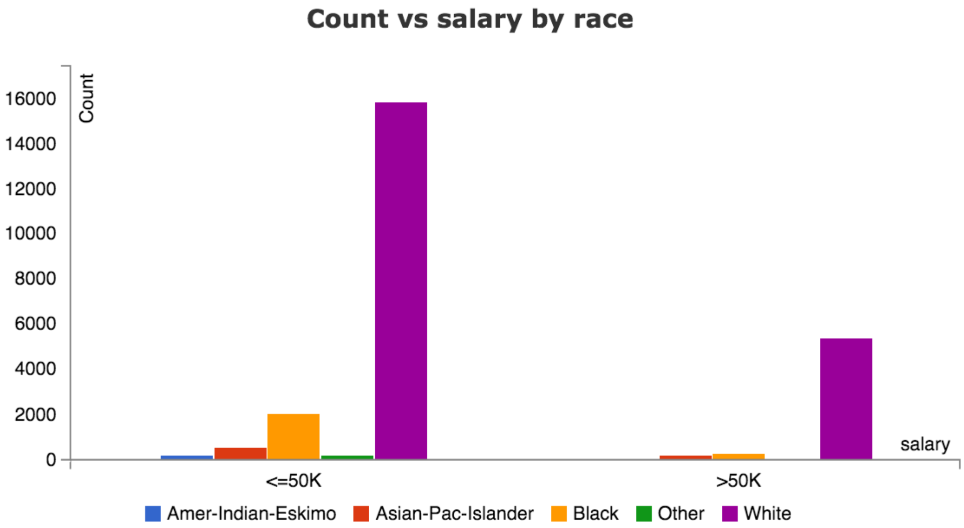


Figure 12: Bar chart of race and salary in the adult dataset

3.3 Initial test

The user wants to study the association between race and salary. In the *Variables selection* tab, he thus selects salary as dependent variable and race as independent variable and launches the test. Both variables are categorical, the sample size is greater than 1000 and race is not a binary variable. Thus eAnalysis selects a chi square test. The results of the test are displayed in the second part of the *Test results* tab. The contingency table summarizes the distribution of the races in the different salary groups. The results of the test (the test statistic and the p value) are presented with a short interpretation. Here the p value of the test is smaller than than the p value set in the parameters. Thus we reject the null hypothesis. There is a significant association between race and salary. At this point, the user can stop. All the necessary elements to report the test (from the test selection process to the test results) are presented in one place. But eAnalysis has also conducted further analysis automatically.

3.4 Post Hoc analysis

First eAnalysis computed the strength of the association between race and salary by using Cramer's V. V is smaller than 0.20 here, which corresponds to a weak association. eAnalysis interprets thus that the association between race and salary is statistically significant but quite weak.
eAnalysis also conducted a test of association for each pair of races. The initial test told the user that race and salary were associated, but brought no information regarding the differences between the races. This analysis allows the user to know which races have a similar distribution of salary (in a statistically significant way) and which do not. The Table 5 presents the results in this case. eAnalysis then groups the races which have similar distribution. Here the White and the Asian-Pac-Islander groups on one hand and that the Amer-Indian-Eskimo, Black and Other groups on the other hand have similar salary distributions.

	p value	**Similar distribution**
(Amer-Indian-Eskimo, Asian-Pac-Islander)	3.492e-06	Yes
(Amer-Indian-Eskimo, Black)	6.408e-01	No
(Amer-Indian-Eskimo, Other)	1.444e-01	No
(Amer-Indian-Eskimo, White)	6.893e-07	Yes
(Asian-Pac-Islander, Black)	2.407e-18	Yes
(Asian-Pac-Islander, Other)	1.728e-10	Yes
(Asian-Pac-Islander, White)	8.652e-01	No
(Black, Other)	2.375e-02	No
(Black, White)	3.124e-44	Yes
(Other, White)	1.036e-09	Yes

Table 5: Assessment of similar distribution between racial groups with regard to salary

Finally eAnalysis computes the adjusted standard residuals for each combination between groups of the dependent variable and groups of the independent variable as represented in Table 6. A higher adjusted standard residual indicates a higher deviation from the null hypothesis. This allows the user to better assess the contribution of each race. Here the White and Black groups have the highest residuals which mean that they have the salary distribution that differ the most from the proportional distribution. However, their residuals have opposite signs for each salary group, which means that they are different in opposite ways. In fact, the White group has an abnormally high proportion and the Black group an abnormally low proportion of members earning more than 50K per year.

	<=50K	**>50K**
Amer-Indian-Eskimo	4.634	-4.634
Asian-Pac-Islander	-1.226	1.226
Black	13.72	-13.72
Other	5.833	-5.833
White	-13.67	13.67

Table 6: Adjusted standard residuals for race and salary

In conclusion, the post hoc analysis has brought many additional insights to the user. The initial test only told him that there was an association between race and salary. But he now knows that this association is pretty weak, has recognized two groups of races with similar distributions of salary and has assessed that the Black and White groups were the most different from the proportional distribution with the White group being abnormally rich and the Black group abnormally poor.

3.5 Confounders analysis

eAnalysis now tries to detect and account for confounding variables that could skew the relationship between race and salary and lead the user to make erroneous conclusions. The first step is to detect which variables could be potential confounders. As per the results presented in Table 7, the final weight (fnlwgt) is not associated with salary and can be eliminated for the search for confounders.

All the other variables are potential confounders. For example, age is associated

	salary	**race**	**Potential confounder**
age	Associated	Associated	Yes
capital_gain	Associated	Associated	Yes
capital_loss	Associated	Associated	Yes
education	Associated	Associated	Yes
education_num	Associated	Associated	Yes
fnlwgt	Not associated	Associated	No
hours_per_week	Associated	Associated	Yes
marital_status	Associated	Associated	Yes
native_country	Associated	Associated	Yes
occupation	Associated	Associated	Yes
relationship	Associated	Associated	Yes
sex	Associated	Associated	Yes
workclass	Associated	Associated	Yes

Table 7: Assessment of potential confounders

with both salary and race. The association between age and salary is quite intuitive as salaries increase with the work experience which is itself correlated with age. However, the association between age and race is more surprising. To better understand these associations, the user can use the data exploration tool.

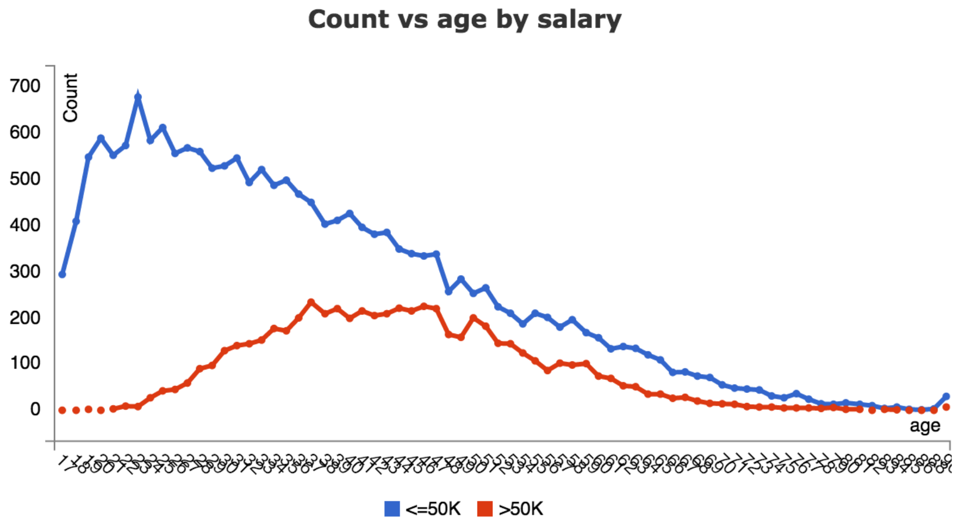


Figure 13: Line chart of age and salary in the adult dataset

The Figure 13 is a graph representing the association between age and salary. The user can see that, indeed, the persons earning more than 50K are statistically older than those earning less than 50K. The user can also obtain the frequency of each age in a particular race, as represented in Figure 14. We displayed only the age frequency for White, Black and Asian-Pac-Islander in order to obtain a better visualisation. By hovering the mouse on each point, the precise value will be displayed. Here we can see that, for Asian-Pac-Islander, 4.5% of the sample are 35 years old. The data can also be exported as a CSV table and be used to adjust the statistical models for further confounders analysis.

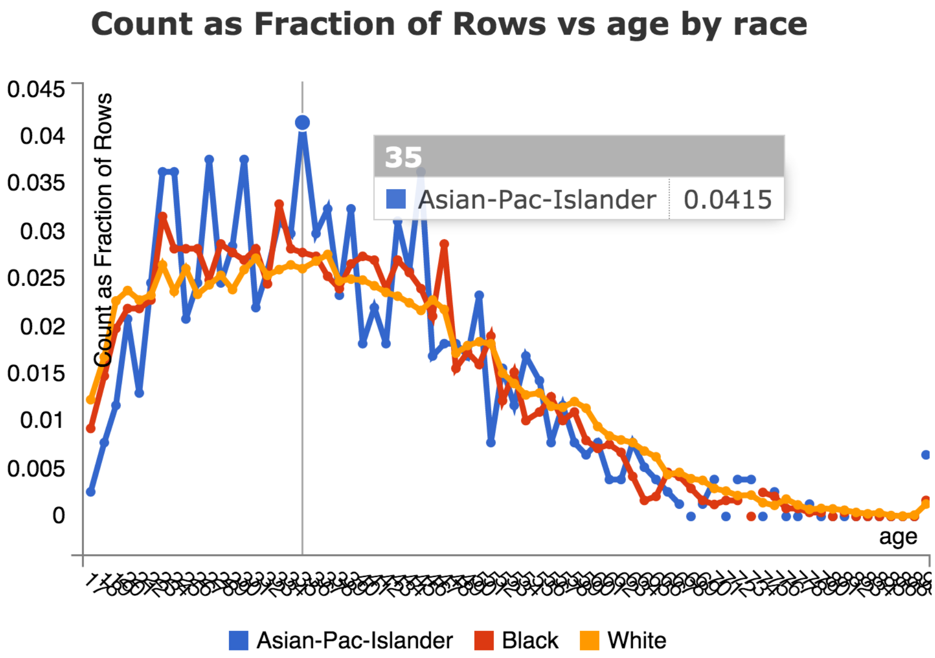


Figure 14: Line chart of the frequency of age for the races White, Black and Asian-Pac-Islander in the adult dataset

The second part of the *Confounders analysis* tab allows the user to study in more details the effect of each potential confounder. For example, the user does a stratified analysis with education as a potential confounder. eAnalysis conducts a test of association by filtering each level of education. There is an association between race and salary for most education levels (although the results must be considered carefully because the sample size might be too small) but not for assoc-acdm.

4 Limitations and extensions

4.1 Limitations and caveats

In this section, we will address the limitations of eAnalysis and the reasons underlying these limitations.

Logistic regression The first limitation is that eAnalysis does not allow to study a hypothesis where the dependent variable is categorical and the independent variable is numerical. To conduct such an analysis would necessitate to implement logistic regression in Python. But logistic regression is not available in scipy. In general, python suffers from a lack of implementation of statistical tests compared to R, which as, in turn, limited the tests available for eAnalysis. We had considered adding the implementation for such tests ourselves. However, implementing a statistical test correctly would need a lot of work to ensure that the implementation is correct (numerous test cases for example) and is out of the scope of this project. As an imperfect solution, the user can get an idea of the relationship between the two variables by selecting the numerical variable as the dependent variable and the categorical variable as the independent variable. It will not be equivalent to the study initially intended but can still bring valuable insights.

Adjustment models for confounder analysis At the moment, eAnalysis offers a module for confounder analysis but this module only offers an analysis by stratification. Another possible way to conduct a confounder analysis would be to adjust the statistical model to consider the influence of the confounder. But this approach is quite complicated to implement. In fact, the adjustment would depend on the type of the variable. Typically, if the dependent variable is numerical, you would do a multi-linear regression, and if the dependent variable is categorical, you would do a logistic regression. However the logistic regression is not implemented in eAnalysis as stated in the above paragraph. Furthermore the standard way of controlling for confounders by using a logistic regression is limited to a binary dependent variable [46]. There have been recent developments to extend this method to multi level categorical dependent variables [47]. But even these techniques require the independent variable to be binary.

Statistical hypothesis testing To select the most appropriate statistical test, eAnalysis assesses the validity of the test underlying hypotheses. To assess this validity, a statistical test is often used. It produces a p value which is then compared to the p value entered as a global parameter and the test confirms or not the validity of the hypothesis. The problem here is that a very small change of the p value can completely change the result of the test. This is a classical criticism against frequentist statistics [48] and this is ultimately unavoidable since eAnalysis needs to make a decision. However, this criticism against the sensitivity to the p value is even more valid when assessing a test hypothesis since studies have shown that statistical tests can be quite insensitive to moderate violations of the hypotheses [49, 50]. In a normal statistical study, the statistician can typically draw a graph to visualize the data and then make a decision if the hypothesis is verified. This eye test can thus complement the statistical assessment. One way to implement this behaviour in eAnalysis would be to plot the data and ask the user for its assessment. This would allow more flexibility to the statistical process and ensure that the user is better implicated.

4.2 Extensions

eAnalysis could be further extended in the future in order to allow a better statistical analysis.

First we could add more statistical tests. On one hand, we could add new test classes to add more possibilities in the choice of variables to analyze, such as ANCOVA to be able to deal with a covariate, or the logistic regression to deal with a categorical dependent variable. On the other hand, we could add tests to existing test classes with different hypotheses, which would allow us to better select an appropriate test. For example, we could add the Mann–Whitney U test, a non parametric equivalent of the T test, which albeit less powerful than the T test, does not need a normal distribution of the data. However these tests are not implemented in scipy at the moment and adding them would require to implement them directly which can be very difficult. Alternatively they could be ported from their implementation in R.

Another possible extension would be adding support for more data formats for the input. At the moment, only CSV files are supported as it is the most common format. But, we could leverage pandas capabilities to support JSON and XML files.

We could also add the possibility of applying filters to the data directly in eAnalysis. Even if the performance should not be a problem, it would allow the user more flexibility in the data analysis. For example, for the US Census dataset, the user could choose to study only the white males or the people older than fifty years.

eAnalysis could also integrate more graphical visualizations in the workflow, such as histograms or scatter plots. These graphs could be used to help the user explore the data in a more intuitive way. Graphs could also be used in the hypothesis testing phase. At the moment, all the hypothesis tests are done by the program. However, it is a classical technique in statistics to plot data and verify hypothesis by sight. eAnalysis could then present a graph to the user and ask him to assess if an hypothesis is verified. This would make the software more interactive and allow the user to be more implicated in the study and thus help him to better understand the statistical process.

In Section 4.1, we briefly presented a criticism of frequentist statistics, namely that they attempt to answer by yes or no to a complex problem, and how this criticism was relevant to eAnalysis. An ambitious way to address this issue would be to rewrite eAnalysis completely to use a Bayesian approach instead of frequentist statistics. Some studies have been conducted to explore this novel approach with promising results [51].

5 Related work

In this section, we will present different tools that address the same problem as eAnalysis and compare their specifications.

There exist multiple tools to do statistical data analysis. Systems such as R [52] or SPSS [53] are very powerful tools to do statistical analysis. They exist since years and are supported by an important community and a big company (IBM) respectively. Both offer a wide variety of statistical tests. However, because in part of their power, they are quite complicated to use and are primarily useful for people with a strong training in statistics. They do not guide the user during a statistical analysis or test statistical assumptions automatically. Furthermore they do not generate additional analysis for the user. In fact, these systems can be so complicated that there exist entire tools to help users to use R or SPSS. For example, Laerd Statistics [54] offer a complete set of interactive guides to help non experts to use SPSS. Laerd offers an interactive experience where the user is directed on a certain path according to the setting of the study. For example, Laerd makes sure that the user verify the statistical hypotheses underlying a test. This approach conducts to a very natural workflow and Laerd was an inspiration in designing eAnalysis. However, eAnalysis goes further by directly integrating the statistical tests in the workflow. This gives the user an even more natural workflow but the risk is that the user no longer fully understands the statistical process. For this reason, eAnalysis gives detailed explanations of its process at each step. Furthermore, Laerd Statistics is a paying service for a proprietary system while eAnalysis is completely free and built exclusively on open source software. We should also mention Redhyte [55], a tool which tries to address the same problem as eAnalysis. It also offers the verification of statistical hypotheses, but focuses more on discovering new hypotheses, using data mining techniques. The workflow is also different: whereas eAnalysis assesses the statistical hypotheses to select the most appropriate test among several tests, Redhyte select a test first and then checks for the validity of the hypotheses. Redhyte also only allows the study of binary variables whereas eAnalysis can study categorical variables of any level.

Conclusion

In this report, we have presented eAnalysis, a web application that aims to automate the process of statistical analysis. eAnalysis is designed to offer a familiar experience to its user by reproducing classical workflows while helping them to be more efficient and rigorous. Through eAnalysis, users can explore the data, conduct valid statistical tests and discover new insights. eAnalysis can be used alone or in conjunction with other statistical packages. The results obtained can conveniently be used in a report or shared with others thanks to detailed explanations at each step of the process.

eAnalysis serves also as a demonstration that a stack equivalent to Shiny in R can be developed in Python, from the statistical process in the back-end to the graphical interface and the web hosting on the front-end. It is still much more difficult to set up that a Shiny app but we believe that, thanks to the work of the open source community, it will become easier with time. eAnalysis is an open source project, built using only open and free software, and external contributions are welcome . Its code has been purposely designed with a modular structure in order to be easy to understand and to extend.

You can find a working version of eAnalysis at `128.199.231.116:8000/notebooks/easy_analysis.ipynb`. The source code is hosted on BitBucket at `https://bitbucket.org/emilebres/eanalysis/src`.

References

[1] SINTEF *Big Data, for better or worse: 90% of world's data generated over last two years.* ScienceDaily. www.sciencedaily.com/releases/2013/05/130522085217.htm

[2] Josh James *Data Never Sleeps 3.0* Domosphere. 2015. https://www.domo.com/blog/2015/08/data-never-sleeps-3-0/

[3] FABERNOVEL *GAFAnomics: A new economy wrapping the planet* Medium. 26 January 2016. https://medium.com/inside-gafanomics/gafanomics-a-new-economy-wrapping-the-planet-part-1-31be978c00f2

[4] Viktor Mayer-Schönberger, Kenneth Cukier *Big Data: A Revolution That Will Transform How We Live, Work, and Think* Eamon Dolan/Mariner Books, 2014.

[5] Darell Huff *How to lie with statistics* Paperback edition, 1954.

[6] Simpson, E. H., *The interpretation of interaction in contingency tables.* Journal of Royal Statistical Society B 13, 238–241 (1951)

[7] Bickel, P., Hammel, E. and O'connell, J. *Sex bias in graduate admissions: Data from Berkeley.* Science 187, 398-404. 1975.

[8] Davenas, E. and Beauvais, F. and Amara, J. and Oberbaum, M. and Robinzon, B. and Miadonnai, A. and Tedeschi, A. and Pomeranz, B. and Fortner, P. and Belon, P. and Sainte-Laudy, J. and Poitevin, B. and Benveniste, J., *Human basophil degranulation triggered by very dilute antiserum against IgE*, Nature, Jun. 1988 V. 333, pages 816–818

[9] Oliver Kirchkamp Workflow of statistical data Analysis 2009 http://www.kirchkamp.de/oekonometrie/pdf/wf-screen2.pdf

[10] J. Scott Long The Workflow of Data Analysis Using Stata, Stata Press, 2009

[11] McDonald, J.H. *Handbook of Biological Statistics* (3rd ed.). Sparky House Publishing, Baltimore, Maryland.

[12] Amir Gandomi, Murtaza Haider *Beyond the hype: Big data concepts, methods, and analytics* Ted Rogers School of Management, Ryerson University, Toronto, Ontario M5B 2K3, Canada 2014. http://www.sciencedirect.com/science/article/pii/S0268401214001066

[13] Thomas A. Glass, *Understanding Obesity from Epigenetics to Communities* John Hopkins University. http://www.globalobesity.org/our-projects/project-1/ Retrieved April 2016.

[14] Tukey, John (1977), *Exploratory Data Analysis*, Addison-Wesley.

[15] adult dataset, UCI Machine learning repository. http://archive.ics.uci.edu/ml/datasets/Adult

[16] Du Prel, Jean-Baptist et al. *"Choosing Statistical Tests: Part 12 of a Series on Evaluation of Scientific Publications."* Deutsches Ärzteblatt International 107.19 (2010): 343–348. PMC. Web. 9 Apr. 2016.

[17] Marenco, Anne. *When to use what test* California State University, Northridge `http://www.csun.edu/~amarenco/Fcs%2520682/When%2520to%2520use%2520what%2520test.pdf` Retrieved September 2015.

[18] Fisher, R. A. (1922). *On the interpretation of $\chi 2$ from contingency tables, and the calculation of P.* Journal of the Royal Statistical Society 85 (1): 87–94. doi:10.2307/2340521. JSTOR 2340521.

[19] Pearson, Karl (1900). "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling" (PDF). Philosophical Magazine Series 5 50 (302): 157–175. doi:10.1080/14786440009463897.

[20] Mehta, C. R. and Patel, N. R. (1986) Algorithm 643. FEXACT: A Fortran subroutine for Fisher's exact test on unordered r*c contingency tables. ACM Transactions on Mathematical Software, 12, 154–161.

[21] Cramér, Harald. 1946. Mathematical Methods of Statistics. Princeton: Princeton University Press, p282. ISBN 0-691-08004-6

[22] Dunn, Olive Jean (1961). "Multiple Comparisons Among Means" (PDF). Journal of the American Statistical Association 56 (293): 52–64. doi:10.1080/01621459.1961.10482090.

[23] Bron, C. and Kerbosch, J. 1973. Algorithm 457: finding all cliques of an undirected graph. Commun. ACM 16, 9 (Sep. 1973), 575-577. `http://portal.acm.org/citation.cfm?doid=362342.362367`

[24] NetworkX, `https://networkx.github.io/`

[25] Agresti, A. (2007). An introduction to categorical data analysis. Hoboken, NJ: Wiley.

[26] Sharpe, Donald (2015). Your Chi-Square Test is Statistically Significant: Now What? Practical Assessment, Research & Evaluation, 20(8). Available online: `http://pareonline.net/getvn.asp?v=20&n=8`

[27] Gosset (William Sealy) aka Student, *The probable error of a mean* Biometrika, Volume 6, Issue 1 (Mar., 1908), 1-25

[28] Welch, B. L. (1947). "The generalization of "Student's" problem when several different population variances are involved". Biometrika 34 (1–2): 28–35. doi:10.1093/biomet/34.1-2.28. MR 19277.

[29] Kruskal; Wallis (1952). "Use of ranks in one-criterion variance analysis". Journal of the American Statistical Association 47 (260): 583–621. doi:10.1080/01621459.1952.10483441 .

[30] Lowry, Richard. “Concepts and Applications of Inferential Statistics”. Chapter 14. http://faculty.vassar.edu/lowry/ch14pt1.html

[31] Stromberg, A. J. (2004). ”Why write statistical software? The case of robust statistical methods”. Journal of Statistical Software.

[32] Shiny, R Studio. http://shiny.rstudio.com/

[33] ”Python” Wikipedia: The Free Encyclopedia. Wikimedia Foundation. 26 March 2016

[34] Python https://www.python.org/

[35] pandas, http://pandas.pydata.org/

[36] scipy, http://docs.scipy.org/doc/scipy/reference/

[37] Jupyter, http://jupyter.org/

[38] Jupyter Notebook extensions, https://github.com/ipython-contrib/IPython-notebook-extensions/

[39] ipywidgets, https://github.com/ipython/ipywidgets

[40] Docker. https://www.docker.com/

[41] jupyter-nb. https://github.com/jupyter/tmpnb

[42] Krutchen, Nicolas. PivotTable.js. https://github.com/nicolaskruchten/pivottable

[43] Cloudlook live benchmark of Digital Ocean droplets. Accessed April 2016. http://www.cloudlook.com/digital-ocean-droplets.

[44] Ronald Fisher *The use of multiple measurements in taxonomic problems* Annals of Eugenics 7 (2): 179–188, 1936.

[45] Cox, D. R. *The regression analysis of binary sequences (with discussion).* Journal of Royal Statistical Society B 20, 215–242 (1958)

[46] Wayne W. LaMorte *Regression Analysis, Controlling for Confounding* Boston University, School of Public Health http://sphweb.bumc.bu.edu/otlt/MPH-Modules/EP/EP713_Regression/EP713_Regression_print.html Consulted in January 2016.

[47] Larsen K, Petersen JH, Budtz-Jørgensen E, et al. *Interpreting parameters in the logistic regression model with random effects.* Biometrics 2000;56:909–14.

[48] Kee-Seng Chia *”Significant-itis” — an obsession with the P-value* Scandinavian Journal of Work, Environment & Health Vol. 23, No. 2 (April 1997), pp. 152-154

[49] Glass, G.V., P.D. Peckham, and J.R. Sanders. *Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance.* Review of Educational Research 42: 237-288. 1972.

[50] Lix, L.M., J.C. Keselman, and H.J. Keselman. *Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test.* Review of Educational Research 66: 579-619. 1996.

[51] Gelman, Andrew. *A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-fit Testing* International Statistical Review, 71-2, Blackwell Publishing Ltd, pages 369–382. 2003.

[52] R Core Team (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/`.

[53] SPSS, IBM. `www.ibm.com/software/analytics/spss`

[54] Laerd Statistics. `https://statistics.laerd.com/`

[55] Toh, W. Z. *Redhyte: Towards a Self-diagnosing, Self-correcting, and Helpful Analytic Platform* National University of Singapore (2015)