

Out of the Ordinary Miner

Data Mining Application

Developers

Arren Antioquia

Arces Talavera

Jet Virtusio

Edmund Gerald Cruz

Rgee Gallega

Supervisor

Arnulfo P. Azcarraga, PhD

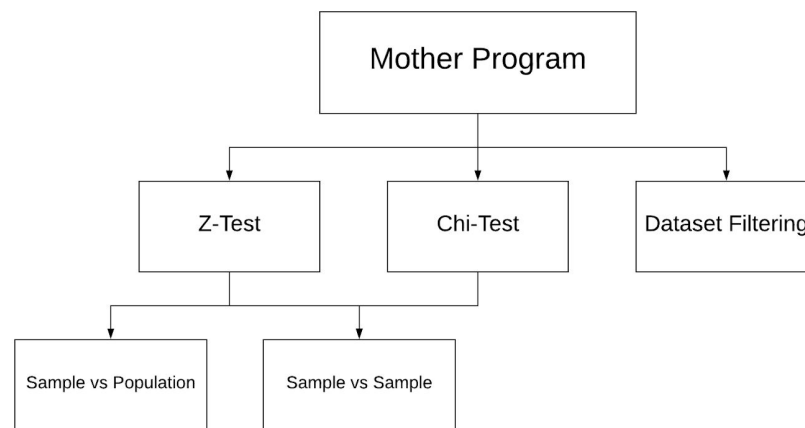
Table of Contents

Table of Contents	2
Introduction	3
System Architecture	3
Procedure	3
Upload your base dataset and variable description	3
Variable Description	4
Variable and Values File	5
Population Dataset	6
Filter your datasets	7
Preview dataset frequency and proportion	8
Z-Test between two samples	9
Chi-test between two samples	10
Z-Test between multiple samples and the population	11

Introduction

The Out of the Ordinary (OOTO) Miner is a data mining application written in Python that analyses two sets of data - filtered from a single base data set - and mines out the features that make them different (or rather, out of the ordinary) from each other. The tests used to compare the two groups of data are chi-test and z-test.

System Architecture



Required Python packages to download:

- Numpy package

Procedure

To start the program, double click on 'Mother.py'.

Upload your base dataset and variable description

- | Feature Marker | Feature Code | Feature Description |
|----------------|--------------|---|
| ^ | c4c2 | I looked for information about the place I want to go visit/eat |
| a | 1 | Never |
| a | 2 | Hardly ever |
| b | 3 | At least every week |
| b | 4 | Daily or almost daily |
| b | 5 | Several times each day |
| b | 6 | Almost all the time |
| -1 | 7 | Prefer not to say |

- **Feature Code** - A short name of your feature. NOTE: It should **always** end with a number. This is displayed as a column header in your dataset.
- **Feature Description** - The descriptive meaning of the feature
- **Response** -
 - **Group** - The class the response belongs to. You can group related responses together. This is needed in order to execute Z-Test properly since it is needed to get the proportion of a common response in all features.

IMPORTANT NOTE: It is required to give **AT LEAST ONE** group, which should be b.

- **Code** - The value of the response. This is what is displayed in your dataset.
- **Description** - The descriptive meaning of the response

Variable and Values File

The variable description can be generated if you provide a values and variable file.

The **variable file** is a text file that contains all of the features and their descriptions. Each feature written should have the following format:

<Feature Code> = "<Feature Description>"

Additionally you can add comments enclosed by `/*` and `*/`. An example of this is shown below:

```
/*Digital ecology*/
d12 = "How were your social media accounts created?"
d13 = "How old were you when you first had your social media account?"
```

The values file is a text file containing all of the features and all of the responses it can have. Each feature should have the following format:

value <Feature Code>

<Group> : <Response Code> = "Response Description"

<Group> : <Response Code> = "Response Description"

....

<Group> : <Response Code> = "Response Description";

An example is shown below:

value b6

a : 1 = "Never"

a : 2 = "Hardly ever"

b : 3 = "At least every month"

b : 4 = "At least every week"

b : 5 = "Daily or almost daily"

b : 6 = "Several times each day"

b : 7 = "Almost all the time"

-1 : 8 = "Prefer not to say";

Note: If you want to invalidate some responses, assign them to -1. These responses will no longer be considered when calculating for frequency, proportion and total of records that answered that feature.

Population Dataset

Your dataset should have no blanks, but it can have invalid response codes that are not in the variable description. Refer to the sample dataset below:

c4c2	c4d	c4f	c4g	c4e	c4h	c4i
1	1	1	1	4	4	1
3	3	1	1	1	1	3
2	2	1	2	2	2	1
2	2	3	1	3	2	1
2	2	1	2	1	2	1
1	2	2	2	2	3	2
1	2	3	2	2	4	3
1	1	1	1	1	1	99
2	1	1	2	1	1	1
1	99	99	1	3	1	99

Filter your datasets

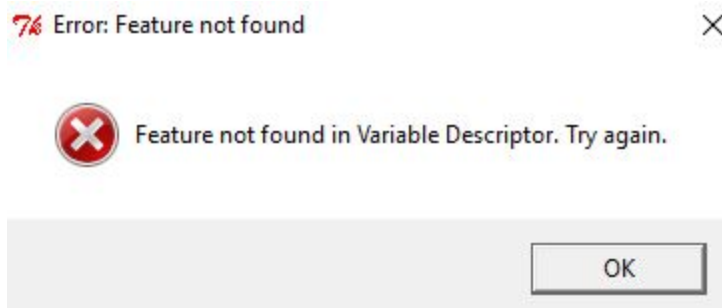
1) Enter feature code

2) Select values of feature to filter by

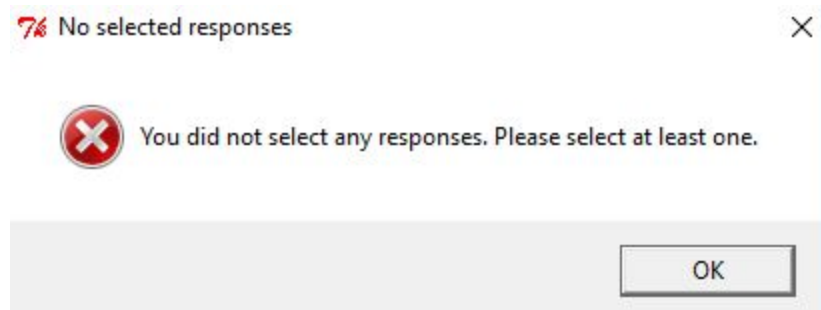
3) Filter. Repeat steps 1-3 until desired dataset is obtained.

Note: You may reset your dataset using the 'Reset Dataset' button.

1. Enter the code of the feature you want to filter by. If the feature is not found in the variable description, the following error message will be displayed:



- The responses of the feature will be displayed. Select at least one value to filter the dataset by; otherwise, the following error message will be displayed:



As you select the values, the number of records of the dataset will be updated, showing how many will be left after you filter.

- Click on the "Filter" button to filter the dataset.
- Repeat steps 1 to 3 until you have your desired dataset. If you want to reset the dataset back to its original state. Click on "Reset Dataset".

After this step, you can do any of the following:

Preview dataset frequency and proportion

- Enter feature code

Enter Code	c7a
0645	34.27% (N) 37.92% (n) a 1 Never
0578	30.71% (N) 33.98% (n) a 2 Hardly ever
0094	04.99% (N) 05.53% (n) b 3 At least every mo
0198	10.52% (N) 11.64% (n) b 4 At least every we
0133	07.07% (N) 07.82% (n) b 5 Daily or almost d
0025	01.33% (N) 01.47% (n) b 6 Several times eac
0028	01.49% (N) 01.65% (n) b 7 Almost all the ti
0012	00.64% (N) 000.0% (n) -1 8 Prefer not to sa

Frequency: 976, Proportion: 57.38%, Total: 1701

- Select values to view frequency and proportion of

1. Enter the code of the feature you want to view the frequency and proportions on. If is not found in the variable description, the following error message will be displayed:

 Error: Feature not found



Feature not found in Variable Descriptor. Try again.

OK

2. Select the responses you want to view the frequencies and proportions of. The columns are organized as:

Frequency | Frequency / N | Frequency / n | Group | Code | Description

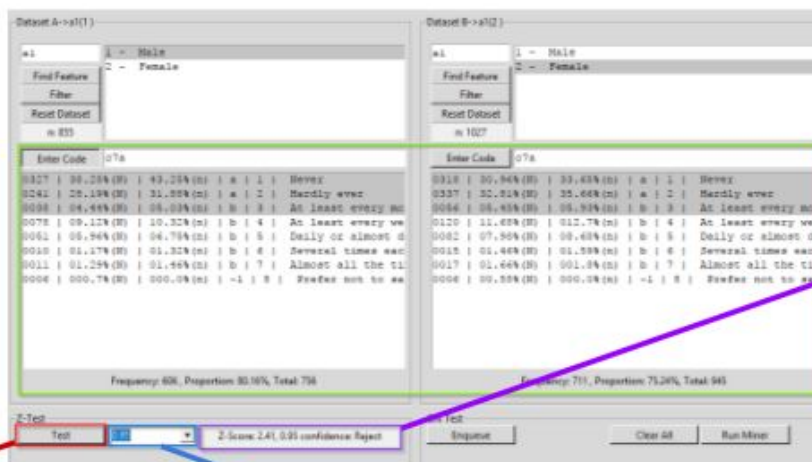
N - The total number of records of that answered that response.

n - The total number of VALIDATED records that answered that response.

Note: A record is considered to be validated if the group of its response at a feature is not -1. All responses of a feature that are not in the variable description are considered to be in group -1.

Z-Test between two samples

- 1) Feature code and selected values for both datasets should be SAME



4) View results

3) Start test

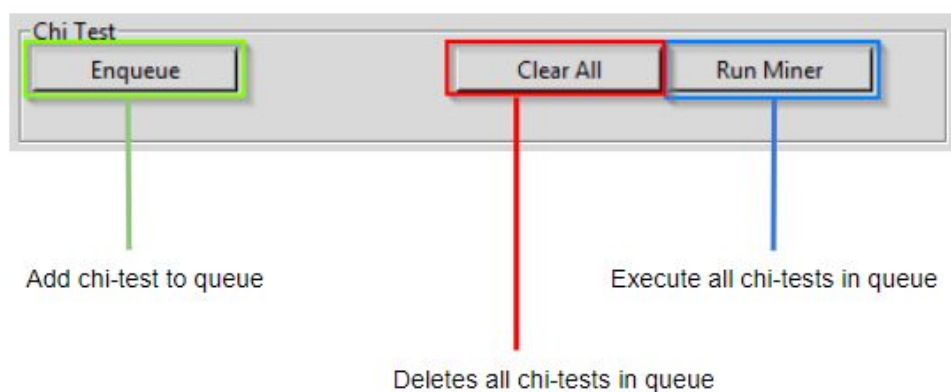
2) Select confidence interval

1. For both sample datasets, enter the feature code and select the responses you want to compare both of them by. Make sure that the feature code and selected responses are the same.
2. Input the confidence interval you desire. This is the threshold that will determine whether the calculate z-score will reject or fail to reject the null hypothesis.

Note: The null hypothesis is that there is no significant difference between the two samples on the given feature)

3. Conduct the Z-Test by clicking on the 'Test' button.
4. View the results. The z-score and whether it is rejected or not will be displayed

Chi-test between two samples



1. Click the 'Enqueue' button once you are ready to set the chi-test between the datasets you filtered. You may reset the datasets if you want to conduct a chi-test between two different ones.
2. Repeat step 1 until you have your desired chi-tests. Click on 'Run Miner' in order to execute all of them. If you want to remove all of them from the queue, click on 'Clear All'.
3. An output file will be made for each test, and should look like the following:

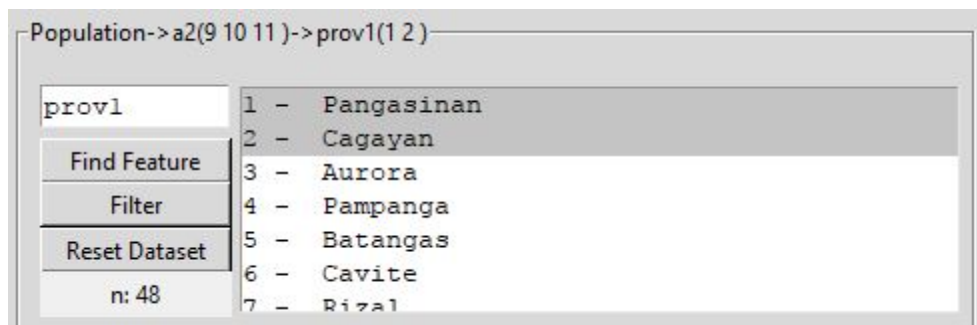
Dataset 1	Dataset 2													
_a1(2).csv	_a1(1).csv													
Feature	Question	Chi	Higher Or	Degrees o	Cut-off	Is signific	N1	N2	P1(a)	P1(b)	P1(etc)	P2(a)	P2(b)	P2(etc)
F16i	I blocked the person from contacting me	8.75901955 +		1	6.635	1	[151.0]	[124.0]	76.82%	23.18%	0.00%	90.32%	9.68%	0.00%
F16h	I changed my privacy/ contact settings	0.863342095 -		1	6.635	0	[151.0]	[124.0]	93.38%	6.62%	0.00%	90.32%	9.68%	0.00%
F16k	Other actions taken	4.034132459 +		1	6.635	0	[151.0]	[124.0]	92.05%	7.95%	0.00%	97.58%	2.42%	0.00%
F16j	I reported the problem online (e.g.; clicked	0.287865841 +		1	6.635	0	[151.0]	[124.0]	90.07%	9.93%	0.00%	91.94%	8.06%	0.00%
F16a	I ignored the problem or hoped the problem	0.126951392 -		1	6.635	0	[151.0]	[124.0]	45.70%	54.30%	0.00%	43.55%	56.45%	0.00%
F16c	I felt a bit guilty about what went wrong	9.370938907 -		1	6.635	1	[151.0]	[124.0]	84.77%	15.23%	0.00%	69.35%	30.65%	0.00%
F16b	I closed the window or app	4.782769727 -		1	6.635	0	[151.0]	[124.0]	86.75%	13.25%	0.00%	76.61%	23.39%	0.00%
F16e	I tried to get back at the other person			1	6.635	0	[151.0]	[124.0]	100.00%	0.00%	0.00%	95.97%	4.03%	0.00%
F16d	I tried to get the other person to leave me a	2.31280821 +		1	6.635	0	[151.0]	[124.0]	84.11%	15.89%	0.00%	90.32%	9.68%	0.00%
F16g	I deleted any messages from the other pers	5.30218191 +		1	6.635	0	[151.0]	[124.0]	74.17%	25.83%	0.00%	85.48%	14.52%	0.00%

- **Higher Or Lower** - Determines which dataset has a higher proportion at Group a.
- **Cut-off** - The chi critical value that is compared with the chi-score to determine the significance of the feature.
- **Is significant** - Whether the the chi-score is greater than the chi critical value or not. If it is, then the chi-score of the feature rejects the null hypothesis.
- **P1(a)** - Proportion of dataset 1 that answered any response with group a.
- **P1(etc)** - Proportion of dataset 1 that answered any response with any of the other groups.
- **N1** - Total number of records in dataset 1 that had a validated response.

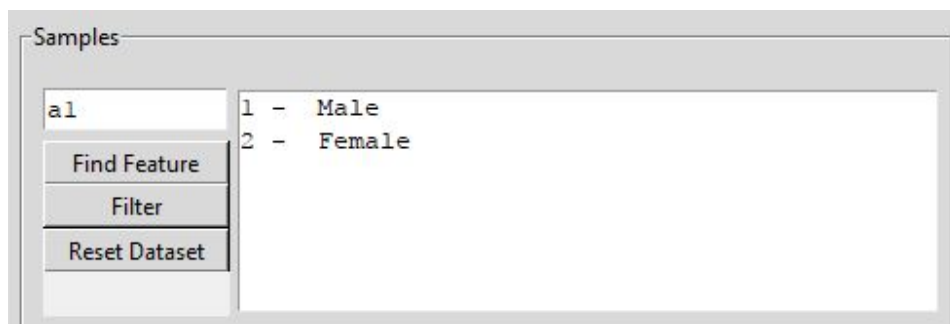
Note: If the chi-score at a feature is **blank**, that means one of the proportions is 0%, making it impossible to calculate the chi-score.

Z-Test between multiple samples and the population

1. Filter the population dataset (if needed).



2. ONLY enter the feature you want to retrieve the samples by.



In the image above, we are getting the samples that are male and female.

3. Select your desired confidence interval and then click 'Test'.

Z-Test Sample Vs Population

Test

0.95
▼

NO DATA

An output file will be made for each comparison between each sample and the population dataset. It should look like the following:

Feature Code	N	F	P	Sample	n	f	p	SE	Z Score	Z Critical	LB	UB	Accept/Reject
a9	825	32	0.038788	1	36	1	0.027778	0.032182	-0.34213	1.96	-0.0353	0.090854	Accept
reg1	72	36	0.5	1	36	36	1	0.083333	6	1.96	0.836667	1.163333	Reject
prov1	72	36	0.5	1	36	36	1	0.083333	6	1.96	0.836667	1.163333	Reject
ur1	855	515	0.602339	1	36	28	0.777778	0.081569	2.150797	1.96	0.617902	0.937653	Reject
a1	855	855	1	1	36	36	1	0	0	1.96	1	1	Accept
a2	570	278	0.487719	1	23	11	0.478261	0.104226	-0.09075	1.96	0.273978	0.682543	Accept
a3a	841	132	0.156956	1	36	2	0.055556	0.060627	-1.67254	1.96	-0.06327	0.174384	Accept
a3b	841	192	0.2283	1	36	8	0.222222	0.069956	-0.08687	1.96	0.085108	0.359336	Accept
a3c	841	816	0.970273	1	36	36	1	0.028305	1.05021	1.96	0.944522	1.055478	Accept
a3d	841	819	0.973841	1	36	36	1	0.026601	0.983378	1.96	0.947861	1.052139	Accept
a3e	841	589	0.700357	1	36	25	0.694444	0.07635	-0.07744	1.96	0.544798	0.844091	Accept
a3f	841	260	0.309156	1	36	15	0.416667	0.077024	1.395806	1.96	0.265699	0.567634	Accept
a3g	841	830	0.98692	1	36	35	0.972222	0.018936	-0.7762	1.96	0.935108	1.009337	Accept

- **N** - The total number of records in the population dataset that had a validated answer to the feature.
- **F** - The number of records in the population dataset in which their answer to the feature is in group b.
- **P** - The proportion of F over N.
- **Sample** - The description of the sample dataset. It is based on the response code of the feature of the population dataset the samples were retrieved by in step 2.
- **n** - The total number of records in the sample dataset that had a validated answer to the feature.
- **f** - The number of records in the sample dataset in which their answer to the feature is in group b.
- **p** - The proportion of f over n.
- **SE** - The standard error of the sample to the dataset on the feature.
- **Z-Score** - The z-score calculated.
- **Z-Critical** - The z-critical value compared to the z-score. Based on the confidence interval selected.

- **LB** - Lower bound calculated based on the z-critical value, p and standard error of the sample.
- **UB** - Upper bound calculated based on the z-critical value, p and standard error of the sample.
- **Accept/Reject** - Reject if P is not within LB and UB. Accept if it is.

Note: If there is a feature where the values are 0 (as shown below), some flags may be raised.

Feature Code	N	F	P	Sample	n	f	p	SE	Z Score	Z Critical Value	LB	UB	Accept/Reject		
b1		0	0	0	1	0	0	0	0	1.96	0	0	Accept	Sample total is 0	Population total is 0

Flags:

- Sample/Population total is 0 - The reason why this may happen is because there is no set of responses for that feature in the variable description.