Statistical methods for the detection of falsified data by interviewers and application survey data in Africa

Souleymane Diakité ENSAE-Senegal, e-mail : souleymanediakite2@yahoo.fr

Abstract

Data quality has a significant impact on the results of analyzes. The concern for quality is all the more justified, if those responsible for the collection are not the professional trade. According to the director of the Institute of Statistics of Mali, 80% of people working in the field of statistics in Mali are not statisticians.

In this work, we applied several methods to detect falsified data. Including law Benford, hierarchical and mixed ascending classification or discriminate analysis. Indicators used: the percentage of extreme values, the percentage of missing values, the percentage jump so the percentage of modality "Other." The results show that the classification seems to be better compared to the application of Benford's law or discriminate analysis. Also the best indicators for the detection of falsified data are ratios of extreme values and missing values. These ratios are much lower in the falsifiers.

Key Words: Falsifiers, Interviewers, Benford laws, classification, discriminant analysis.

1. Introduction

Data quality is one of the main concerns of users. Data quality can be affected by different ways, including, among other poor design media collections, the bad answers provided by the respondent or forgery by the interviewer it is the latter that concerns us in this study. Several authors including Schreiner, Pennie, and Newbrough (1988), Schräpler and Wagner (2003) and more recently by Sebastian Bredl, Kötschau Kerstin and Peter Winker (2012).

The work is applied several methods for detecting counterfeiters then compare the results. Thus we will apply a set of methods including Benford's law, the hierarchical classification after factor analysis, the joint classification and discriminate analysis. These methods allow interviewers to characterize risks from some indicators defined on the characteristics of the responses (extreme responses, missing values, the number of hops, the time of filling the questionnaires, the number of completed questionnaires ...)

2. Results of statistical method for detecting tampered data.

2.1. The Benford's law

Benford notes that the probability of the first non-zero number of digits can be described by the following law: $P(d) = Log_{10}(1 + \frac{1}{d})$ for $(d = 1,2,3,4,5,6,7,8,9)$. We have: $\sum_{d=1}^{9} P(d) = 1$

This law is widely used especially in the field of detection of financial fraud. It was used by Swanson and al. (2003) to show that the distributions of the first digits of numbers in the "Consumer Expenditure Survey of the United States" followed the Benford distribution. The idea is that a significant difference in the distribution of first digits of an investigator with the i Benford indicate a risk of falsification of figures that investigator. This difference can be measured with several indicators including the chi-square distance.

$$\chi_i^2 = n_i \sum_{d=1}^{9} \frac{(X_{id} - X_{ad})^2}{X_{ad}}$$

n_i : the total number of digits in the first survey of individual i

X_{id} : the proportion of the first digit in the questionnaires individual i

X_{ad} : the proportion of first digit according to Benford's law.

A value χ_i^2 too high indicates that the interviewer i is an "investigator at risk."

The data come from a survey conducted by the National Superior School of Statistics and Economic Analysis (ENSAE) Senegal's Cyber Cafes and users. The survey was conducted by students Works Engineer Statistics (ITS) in the second and third years of training we denote respectively by T1, .., T11 and F1, ..., F22

It turned out that some engineering students (especially those of the third year) who already had to make inquiries in the past have not been on the field to meet users cyber cafes and generated data. The objective of the work will be to analyze the data with a view to know the risk of tampering with the interviewers . We will in the first instance, from the methodologies presented above regarding Benford's law, analyze the data quality.

The results showed that about 32 interviewers not involved the study only 10 meet the criteria for Benford's law for a risk probability of 5% corresponding to a chi-square $\chi_i^2 = 15{,}4$. These interviewers are ten (F3, F6, F15, F16, F18, F19, F20, T3, T5, T8) represent only 31.25% of total interviewers . This low percentage allows us to say that in this context the application of Benford's law to detect fraudulent data gives a rather mixed results.

Another major limitation of Benford's law is that it is only usable on quantitative variables questionnaires. Or falsification concern quantitative and qualitative variables as well.

2.2. The methods of factor analysis

Two methods of factor analysis can be used. This is the automatic classification and discriminate analysis. The latter requires a priori knowledge of forgers. The main idea is to use a number of indicators to highlight the falsifiers and make a classification as a result of a factor analysis of these indicators.

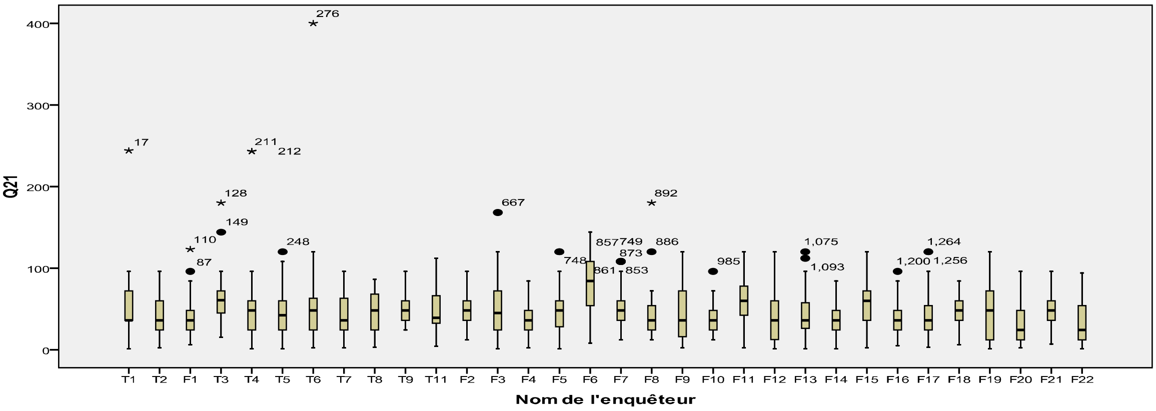
In the literature, Schafer et al. (2005) assumes that observed fewer missing values in the falsifiers. So they tend to respond to all questions. The first indicator obtained as a result of this situation is the "***partial non-response rate.***" Defined as the ratio of missing values on the total number of questions. In addition to Schafer et al. (2005) other authors such as English and Porras (2004) finds that the falsifiers choose less extreme answers to common questions, that is to say, the answers that seem more likely. Based on this observation, we can define a second indicator as the "ratio of extreme answers for measure questions" measured by the ratio of the number of extreme responses to total responses. Another observation is that counterfeiters tend not to choose the modality "***other specify***" for the semi-open questions, we defined a third indicator as the "***ratio other modality to specify***" in relation to all questions. This ratio should be low to the falsifiers. The fourth indicator is the fact that counterfeiters tend to choose the "no" response to the screening questions for failing to answer the questions below. We define a fourth indicator as the ***"ratio of no answers or jumps"*** to the filter questions which should be high in the falsifiers. Indeed, the choice of response not possible to perform jumps and complete the questionnaire faster.

Other indicators could be included in the analysis as the average time or the number of people interviewed per day Bushery et al. (1999). For this in our example the number of completed questionnaires is fixed in advance and we do not have the means time for interview. This place is also the number of jumps. Indeed, the greater the leaps you take less time to complete the questionnaire.

• Ratio of extreme values for quantitative variables

On this box mustache below, we see the presence of outliers in some interviewers . Note that outliers should be far fewer in falsifiers (F) compared to interviewers were on the ground (T). We consider the values that are beyond the first and last deciles as outliers for quantitative variables. From the box mustache, we can say that some interviewers include (F20, F21, F22, T7, T8, T9, T11 ...) pose risks falsification.

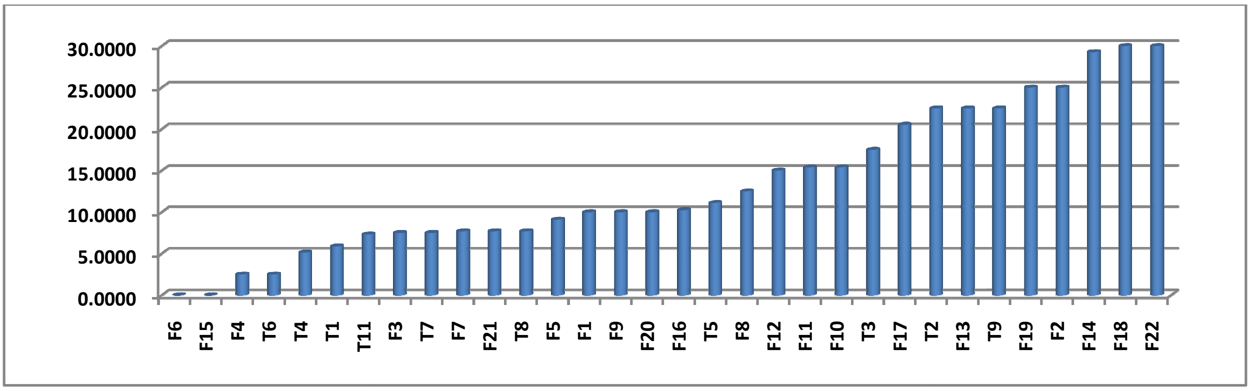
Figure 1: Box plot in time using the Internet according to interviewers



• "Ratio of modality others specify for semi open-ended questions"

As noted above, forgers tend to choose the terms present in the questionnaire. Indeed the choice of modality "other" often requires precision and therefore after further reflection for the forger. On the chart below we have the ratio of other modalities. It is noted that some interviewers have hardly chosen method "Other." These interviewers as (F6 and F15). These interviewers may be suspected of having falsified the data.

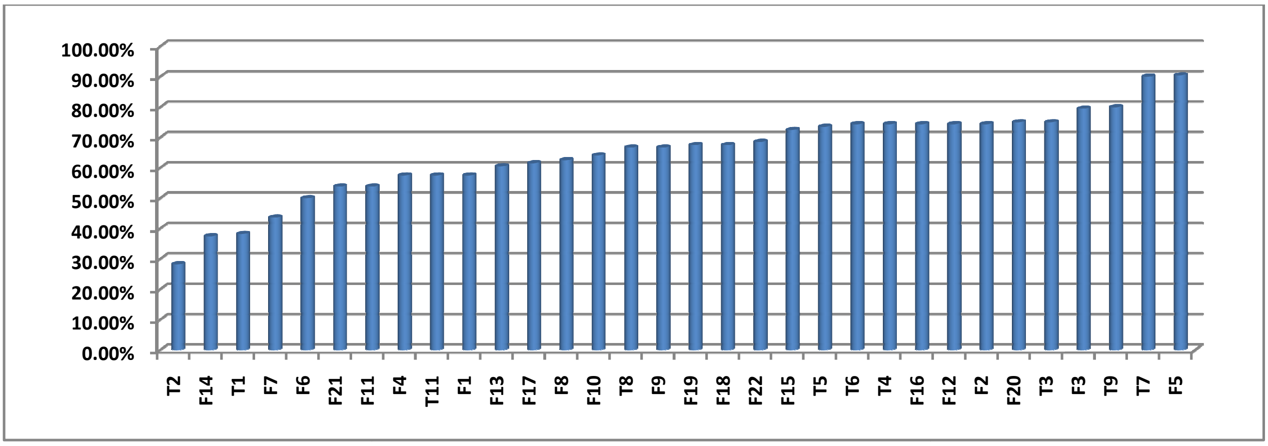
Figure 2: The ratio modality "Other" for the semi open-ended questions



• ratio of answers "no" followed by a jump

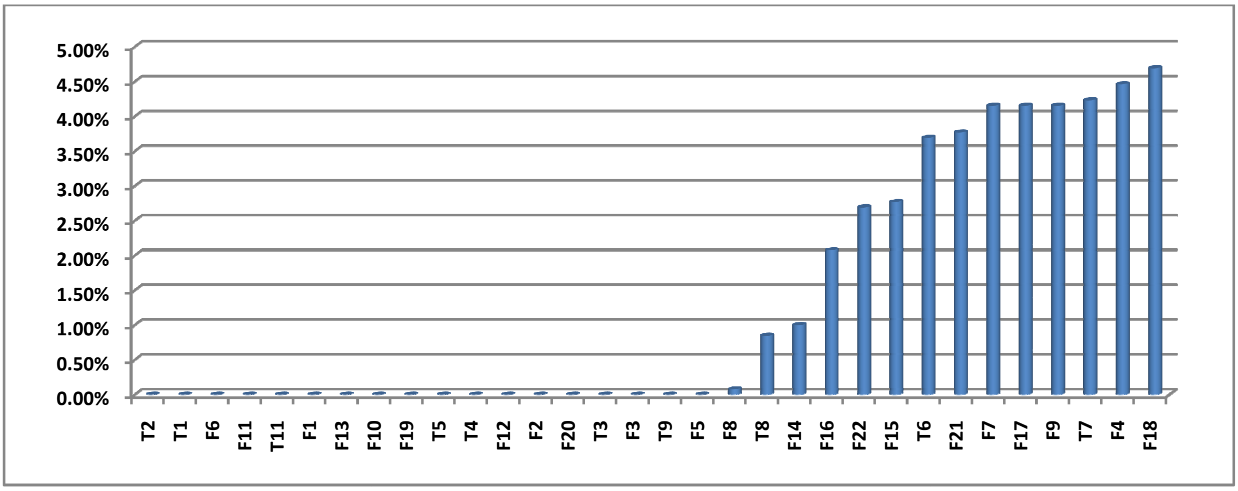
On the chart below, we see the proportions of "no" answers too high for some interviewers. These proportions reaching 90% in some interviewers (T7 and F5). This observation leads to a suspicion of falsification of data from the interviewers.

Figure 3: Proportion of answers "No" followed by jump



• partial non-response rate.
In the chart below, we have the proportion of missing values by interviewers. The absence of missing values for T1 interviewers , T2, F1, T3, T4, T5, T9, T11, F2, F3, F5, F6, F8, F10, F11, F12, F13, F19 and F20 causes a hint of falsification for these interviewers .

Figure 4: Proportion of missing values investigator



• Results of factor analysis
The results of the factor analysis show a number of extreme values opposition to interviewers that the choice of the other modality, while the proportion of missing values appears to be independent of the other two.

• The hierarchical clustering

We performed a hierarchical clustering of factors after factor analysis. We obtain a class composed of twelve individuals who can be called "class falsifiers." Indeed, as shown in the table below, it is characterized by a proportion of missing values lower (an average of 0.25 against 1.34 for the entire population) and much less extreme values (an average of 15.94 against 24.33 for the entire population).

Table 1: indicators that characterize the class of forgers

```
+--------+-------+------------------+------------------+-----------------------------------------------------------------
| V.TEST | PROBA |     MOYENNES     |   ECARTS TYPES   |                     VARIABLES CARACTERISTIQUES
|
|        |       |  CLASSE  GENERALE |  CLASSE  GENERAL  | NUM.LIBELLE                                                  IDEN
|
+--------+-------+------------------+------------------+----------------------------------------------------------------+
|                 CLASSE  1 /        ( POIDS =    12.00      EFFECTIF =   12 )                                          aa1a |
|
|
|        |       |                  |                  |
|
|  -2.61 | 0.005 |     0.25     1.34 |     0.60     1.79 |   5.Missing                                                   C6
|
|  -2.93 | 0.002 |    15.94    24.33 |     6.28    12.35 |   2.Extreme                                                   C3
|
+--------+-------+------------------+------------------+----------------------------------------------------------------+
```

On the individuals who composed, we find both students in the second year than the third year. Individuals who compose it are F1, F2, F10, F8, F13, F16, T1, T3, T4, T8, T9, T11.

Table 2: individuals suspected false after the classification

```
------------------------------------------------------------------------
|RK | DISTANCE  | IDENT. ||RK | DISTANCE  | IDENT. ||RK | DISTANCE  | IDENT. |
+---+-----------+--------++---+-----------+--------++---+-----------+--------+
|  1|    0.07920|F10     ||  2|    0.29510|F8      ||  3|    0.59224|T8      |
|  4|    0.70576|T3      ||  5|    1.02946|F1      ||  6|    1.24404|F13     |
|  7|    1.30482|T11     ||  8|    1.48205|T4      ||  9|    1.57342|F16     |
| 10|    2.65791|F2      || 11|    3.34308|T9      || 12|    4.46437|T1      |
+---+-----------+--------++---+-----------+--------++---+-----------+--------+
```

• Mixed Classification:

To analyze the robustness of the hierarchical clustering obtained, we took the classification using the method of mixed classification. Indeed Hierarchical Clustering has the unseemly not be a global optimum in the sense that the partition constructed at a given level depends on the score obtained in the previous step. The idea of mixed classification is to try to get as close as possible to the optimal classification if it is using the joint use of the Hierarchical Clustering and Classification of Mobile centers. The results give us a class of "falsifier" characterized by only a small proportion of extreme values (15.73% against 24.33% for the total population) in contrast to the hierarchical classification where we had a class of "falsifiers "characterized by a low proportion of missing values and outliers. The class is composed of 15 individuals from whom we have 12 individuals in the Upward classification (F1, F2, F10, F8, F13, F16, T1, T3, T4, T8, T9, T11) plus three individuals who are F14, F18, T2. Ultimately we can consider as falsifiers of 12 individuals confirmed by the Joint method CAH. Indeed, these individuals have statistically lower than those of other interviewers missing and extreme values. NB: Some of these interviewers is found to have cheated at the end of the investigation it is particularly interviewers F2, F10, F13, T9, F16 ...

Table 3: individuals suspected of tampering after the mixed classification

```
EFFECTIF:   15
------------------------------------------------------------------------------
|RK | DISTANCE  | IDENT. ||RK | DISTANCE  | IDENT. ||RK | DISTANCE  | IDENT. |
+---+-----------+--------++---+-----------+--------++---+-----------+--------+
|  1|    0.15570|F10     ||  2|    0.48675|F8      ||  3|    0.65321|F13     |
|  4|    1.08539|T3      ||  5|    1.17609|T8      ||  6|    1.25466|F1      |
|  7|    1.72616|T11     ||  8|    2.02974|F16     ||  9|    2.43074|F2      |
| 10|    2.75543|T4      || 11|    3.52119|T9      || 12|    4.17213|T1      |
| 13|    5.19323|F14     || 14|    6.58308|T2      || 15|    8.41758|F18     |
+---+-----------+--------++---+-----------+--------++---+-----------+--------+
```

2.3. Discriminant analysis:

We will determine the variables that best characterize the two classes obtained. By simultaneously taking into account in the analysis. The class variables will be added to the data table, and play the role of variable explained in discriminant analysis from factorial components (variables) and then back to the original variables. The results show that of the 32 individuals, 22 were correctly classified is an error classification rate of 31.25%. In addition to the four variables used in the analysis, only the percentage of extreme values can be well discriminated forgers non falsifiers.

Table 4: calculating the rate of misclassification after discriminant analysis

```
 TABLEAU DE CLASSEMENT
                       POURCENTAGES DES CLASSEMENTS
                       BIEN CLASSES     MAL CLASSES      TOTAL
 GROUPES D'ORIGINE -------------------------------------------------
                  AA_1      14.00            6.00           20.00
                          ( 70.00)        ( 30.00)        (100.00)
                   -------------------------------------------------
                  AA_2       8.00            4.00           12.00
                          ( 66.67)        ( 33.33)        (100.00)
                  -------------------------------------------------
                 TOTAL      22.00           10.00           32.00
                          ( 68.75)        ( 31.25)        (100.00)
```

3. **Conclusion** :

Data quality a central issue in the field of statistics because it affects the results of the empirical analysis. In this work, we applied several methods to detect falsified data. The indicators used in this study are: the percentage of extreme values, the percentage of missing values, the percentage jump so the percentage of modality "Other." The results show that the classification seems to be better compared to the application of Benford's law or discriminant analysis. Also the best indicators for the detection of falsified data are ratios of extreme values and missing values. These ratios are much lower in the falsifiers.

References

- Benford, F. (1938). "The law of anomalous numbers. Proceedings of the American Philosophical Society" 78 (1), 551–572.
- Bredl, S., P. Winker, and K. K¨otschau (2008). "A statistical approach to detect cheating interviewers" ZEU Discussion Paper Nr. 39.
- Sebastian Bredl, Kötschau Kerstin and Peter Winker (2012) “A statistical approach to detect falsification of survey data by interviewers” Statistics Canadal 12-001, 1-13