Submission for the 2013 IAOS Prize for Young Statisticians

The policeman and the statistician

On the quality of the raw data in official statistics

Anton Färnström
Statistician
Unit for Legal Statistics
Crime Prevention Council
Stockholm, Sweden
antonfa@gmail.com

Abstract

As state institutions are implementing increasingly sophisticated computerized systems, tremendous amounts of data are being constantly generated. This has opened up a wide range of possibilities for the production of statistics that are highly relevant to virtually all sectors of society. However, despite a large consensus regarding the importance of such statistics, little research examining the specific quality issues associated to administrative data is available. Here, we report the results of two studies conducted by the Swedish Crime Prevention Council evaluating the quality of the raw data used for the official crime statistics in Sweden. Our analyses revealed that the Police assigned in average 12% of the crime codes incorrectly, with severe consequences for the final statistics. Furthermore, different interpretations and over-usage of a particular code have rendered national statistics on police decisions nearly unusable. Based on our observations regarding the main causes of these errors, we propose a few recommendations likely to be applicable to both users and producers of statistics based on administrative sources. With official statistics being at the backbone of evaluations and strategic decisions, we strongly believe that investigating and continuously improving the quality of the raw data is a must for all the national statistical agencies.

Introduction

With the computerization of the state bureaucracy's allowing for efficient data storage and easy access to information, the interest in developing statistics based on administrative records has reached new heights. Already, registry data is being used extensively for both research purposes (Bakker, 2012), and in the production of official statistics. For example, according to a newly completed governmental investigation, approximately 95% of the Swedish official statistics originate from administrative processes (SOU 2012:83). In this context, developing methods to understand, evaluate and improve the quality of this type of statistics is essential.

Unfortunately, research investigating the characteristics and quality of registry data has not been at par with the rapid increase in the amount of accessible data. What is probably the first comprehensive theoretical book dedicated to registry-based statistics was published as late as 2007 (Wallgren & Wallgren, 2007). A few years later, when Statistica Neerlandica presented their special edition on statistics based on administrative data, they stated in their introductory note that *"Theory of registry based data is scarcely out of the egg"* (Bakker and Daas, 2012).

Furthermore, surprisingly little attention has been given to issues related to the quality of the raw data, and the potential impact that this has on the final statistics. Instead, most of the research so far has focused on other aspects, such as methods to integrate data from different records (Zhang, 2012), the validity of administrative variables (Bakker, 2012) or differences in estimates from surveys and administrative records (Groen, 2012). Exploring these topics helps, but is not sufficient to get a good understanding of the quality issues related to the raw data.

A significant part of the work in the field has been carried out by individual public authorities. A few examples in Sweden include an evaluation of the coding system used by the Educational Registry (SCB, 2006), evaluation of the coding of professions for the Swedish occupational register (SCB, 2007) and evaluation of the quality in the Swedish mortality registry (The National Board of Health and Welfare, 2010). Nevertheless, such studies are often written in other languages than English, and their results are advertised mainly within the institution that conducted them. This makes their findings inaccessible to a broader public, and hinders faster developments in this field.

The Swedish National Council for Crime Prevention (Brå) has commenced a series of quality studies to gain insights into issues associated to the raw data on which the official criminal statistics are based. In this paper, the results from the first two studies are presented. By analyzing the impact of the coding errors on the official criminal statistics, we reveal the importance of investigating and quantifying the extent of such errors when working with registry data. We conclude by formulating a few recommendations that are likely to be applicable in any institution producing or using statistics based on registry data.

Methodology

The Swedish official crime statistics

The crime statistics in Sweden are produced by the Crime Prevention Council (Brå). The data for the statistics are collected from the different authorities in the judicial system: the Police, the Attorneys Chamber, the Courts and the Criminal Care. As part of the information that the Police and the Attorneys Chambers register in their everyday work are the so-called crime codes and decision codes.

The *crime codes* are 4 digit numbers defining the type of crime investigated, and are mostly assigned by the investigators from the Police. The codes cover all possible crimes, following broadly the structure of the law. As an example, a *fraud using internet* and a *fraud using false invoice* have their own unique codes, and they are part of the *Fraud* category. In order to find all cases of *Fraud*, it is necessary to sum up all the individual types.

When the Police are discontinuing an investigation, they need to provide a motivation for this decision. To do this, they have to choose between a number of different *decision codes*, each representing a specific motivation. Altogether, the Police have access to 15 such decision codes, of which one is called *Other*, and has a free text option.

The crime codes and the decision codes are the main building blocks for all the Swedish official statistics regarding reported offences, cleared-up crimes and suspected persons. A correct coding is thus essential for obtaining accurate statistics on these matters. Below we describe the design and methodology of two studies conducted at Brå for investigating the quality of these two code groups. The first of these studies was completed in September 2012, while the second one is ongoing.

Quality study of crime codes

In 2011 Brå carried out a study designed to answer two questions: "*To what degree are the registering policemen making a correct coding?*", and more importantly, *"What are the consequences of incorrect entries for the official statistics?"*

In total, a stratified independent random sample of 1 598 police reports from eight categories of offences was studied. The sample was stratified in order to increase the precision of the estimates in areas where high error rates were suspected. To check whether a certain crime code was correct, the method of *independent verification with adjudication* was utilized (Biemer & Lyberg, 2003). This method has also been employed, with minor variations, in quality studies done by Statistics Sweden (SCB 1978, SCB 2007), and by the National Board of Health and Welfare (National Board of Health and Welfare, 2010). The procedure is outlined in Figure 1.

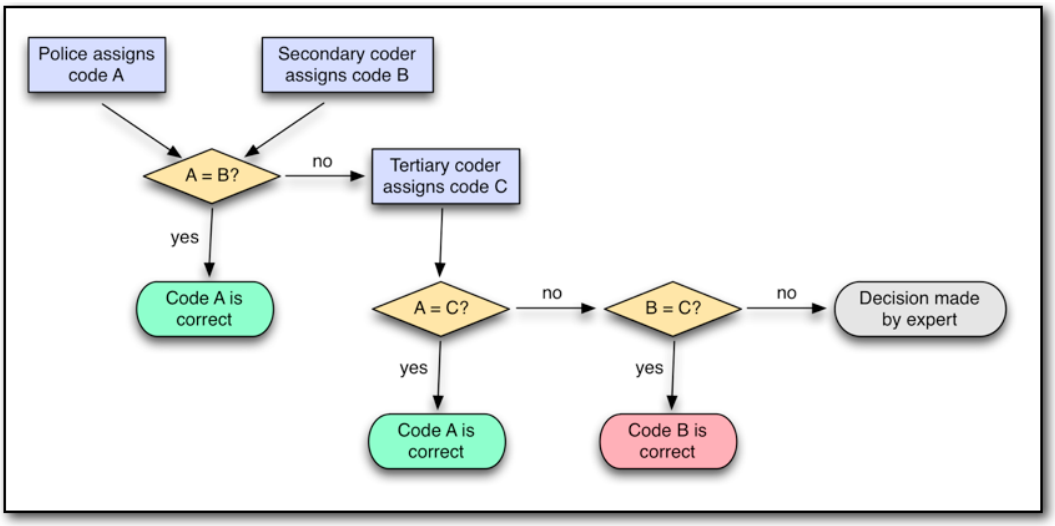

Figure 1. Illustration of *independent verification with adjudication*

By comparing the correct codes with the ones assigned by the Police, a number of indicators describing the quality of the coding were calculated. In our study, emphasis was put on the net error and the misclassification rate, but information regarding the gross error was also provided. These measures and their interpretations are briefly described below, and the details of their calculation are given in the Appendix, at the end of this document.

The *net error* is the difference in the number of offences for a certain crime type before and after recoding. This is an established measure that has been used by Statistics Sweden in several studies (SCB 2007, SCB 2006:4 and SCB 1999:3). It is arguably the most important measure for the users of the official statistics, since it shows the actual effect of the coding errors on the statistical output. The *misclassification ratio* gives the amount of falsely assigned offences divided by the total amount of cases in that crime type. This measure provides an easy interpretable answer to the question: how many faults were in the coding done by the Police? This was the preferred measure of the previous study on coding errors in crime statistics (SCB, 1978). The *gross error rate* is the amount of falsely assigned offences, plus the amount of falsely exempted offences divided by the total number of offences of that type. This gives insight into the total amount of errors that are related to a certain code or category (Holmberg, 2012).

To account for any effects of the random variation of the sampling, two statistical tests were used: a significance test for each individually calculated net error, and a test for the changes in each category as a whole. Detailed descriptions of these tests are given in the Appendix.

Quality study of decision codes

The quality study on the decision codes was originally motivated by an observation that the use of the code called *Other* had greatly increased over the years. Furthermore, we also noted a large variation in the use of this code at the different counties in Sweden (Figure 2).

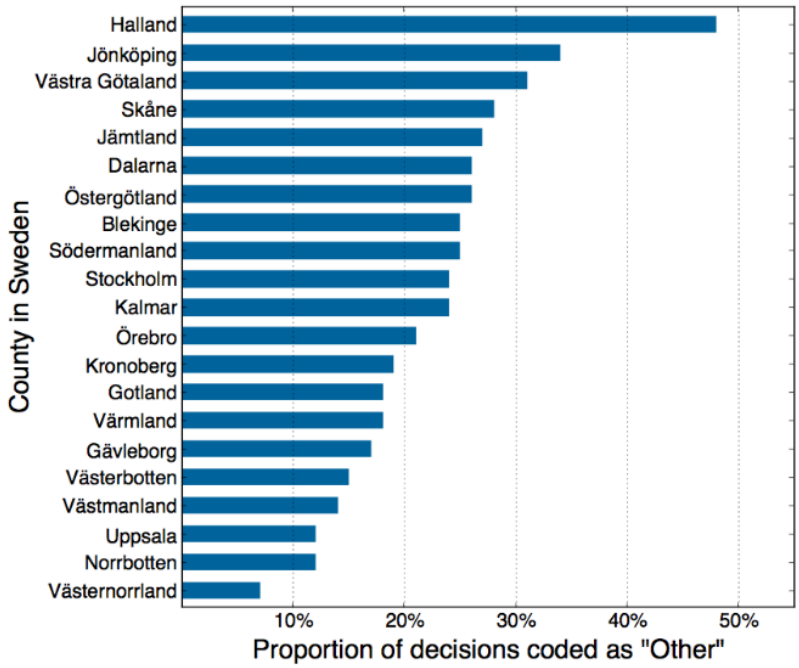

Figure 2. Proportion of decisions coded as *Other*

In the light of these observations, there were two main objectives of this study. First, since the code *Other* includes a free text field where the users fill a motivation, we wanted to find out what those motivations were, and whether the decisions could have been assigned a more specific code. Secondly, we wanted to gain insight into how investigators in different counties interpreted this code.

We included in our study all the decisions from 2011 that were coded as *Other*, a total of approximately 130 000 cases. Since it would have been too labor-intensive to go through all of these free texts manually, we used the statistical program SAS to match the free texts to a library of keywords. The keywords were defined by studying the most commonly used free text motivations, and sorting them to consistent categories. As a quality control, an independent sample of 100 free texts was drawn from each category. If the samples contained unwanted cases, the keywords were changed and the procedure repeated until the categorization was satisfactory.

This method allowed us to classify 81% of all the free text motivations. From the remaining 19%, a simple independent random sample of 1 000 free text records was drawn. These were allocated to categories manually, and then standard statistical methods were used to estimate the distribution of the remaining posts. To summarize, the population was divided into two strata, one comprising 81% of the posts that was automatically mapped, and one comprising 19% of the posts where the distribution in different categories was estimated using random sampling.

A second step of the study was interview-based. We selected three Police authorities from different counties: Stockholm, Uppsala and Jönköping. These Police authorities were chosen due to large differences in the relative amount of usage of the code *Other* (Figure 2). The idea was that this probably reflected differences on the interpretation of how to use the decision codes. At the time when this study was written, three to four police officers had been interviewed in each county.

Results

The quality of the statistics for reported offences

We have evaluated the crime code system used within the Police and the Attorneys Chamber. These codes represent the base of the official Swedish statistics regarding reported offences, cleared-up crimes and suspected persons. The evaluation was carried out by manually examining 1 598 police reports, and assessing in each case whether the description of the crime matched the code assigned by the Police (Figure 1). We selected cases from six different crime categories: *Robbery*, *Fraud*, *Burglary*, *Theft (not burglary)*, *Criminal damage* and *Assault*. To estimate the total amount of incorrectly coded cases, a seventh category consisting of all the codes not included in the previous categories was created.

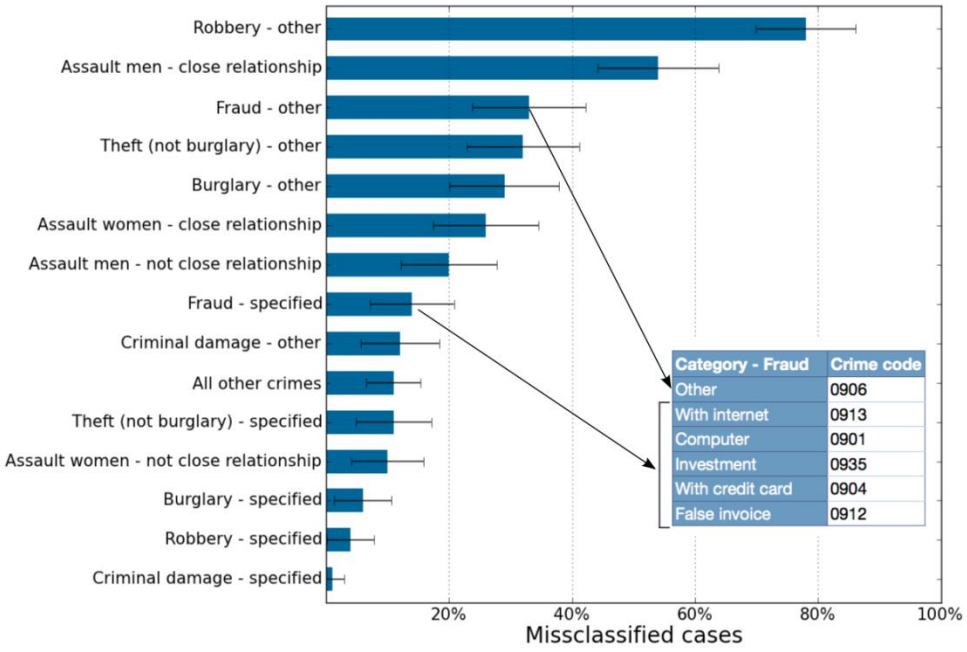

Figure 3. Proportion of incorrectly coded cases

Altogether, we estimated that 12% of all the police reports had an incorrect code assigned. A recurring pattern was the over-utilization of the codes called *Other* in each crime category. To illustrate this, we divided each crime category in two parts: one consisting of all the specified crime sub-categories, and one consisting of the *Other* sub-category (Figure 3).

The crime category *Fraud* is a good example of how the over-utilization of the *Other* codes affected the official statistics (Figure 3, lower box). We estimated that, if the coding was done correctly, the sub-category *Other* should have included approximately 8 400 less cases (–26%). All of these cases should have been assigned a code for one of the other sub-categories of fraud, which thereby should have increased by between 15-20% each. However, it is important to emphasize that none of the cases studied for *Fraud* should have been coded as anything else than fraud. In general, this applied for all of the studied crime categories. The only exceptions were *Burglary* and *Theft (not burglary)*. For these cases, our analyses showed that, apart from the coding errors within each category, there is also a problem with

the delimitation between the two categories. To sum up, the quality of coding of crime codes are reliable at the category level, while highly unreliable for sub-categories.

The registering of relationship status for assaults

Assaults within close relationships are a prioritized field for the Swedish Police, and a topic that is often mentioned in debates about police work in Sweden. Thus, whether an assault can be classified as *in close relationship* or not is of very high public interest. Our study examines for the first time the quality of the registrations of these cases.

Our results have revealed severe problems in the understanding of the definition for *close relationship*. Approximately every second offence registered as assault against men in a *close relationship* was incorrect. Similarly, a quarter of the cases reported as assault against women in *close relationship* should have been coded with a different relationship status. When correcting for these errors, the final statistics changed dramatically (Figure 4).

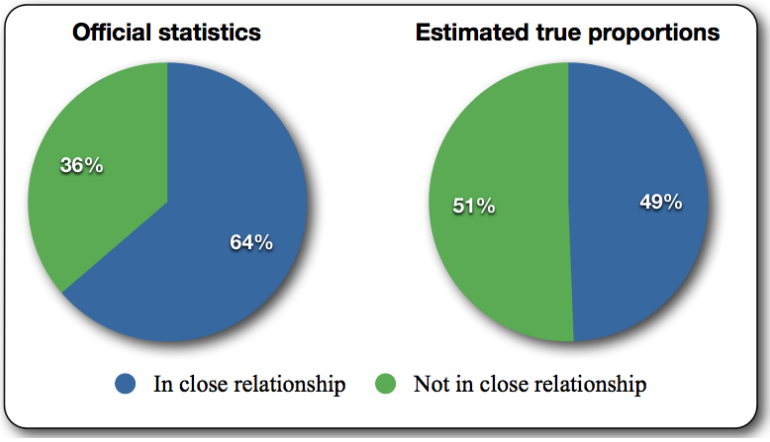

Figure 4. Proportions of assaults in close relationship, before and after recoding

When examining the reasons for those large coding errors, we found a discrepancy between the official definition, and how the police officers appear to have understood the concept of close relationship. In the official definition, the concept is very restrictive[1], going often against the intuition. As an example, most people would consider that sibling relations are close, but this is not the case according to the official definition.

The registration of decisions to cease police investigations

When the Police decide to cease an investigation, they register their decision by choosing among 15 different codes, each corresponding to a specific reason. One of these codes is called *Other*, and instead of a pre-specified description, includes a free text where the registering officer writes the reason for ceasing that investigation. We have examined the content of these free text motivations, and found that approximately 65 000 decisions could have been represented with another code. This translates to a 48% decrease of the number of cases with this code. Consequently, the statistics regarding other specified codes changed substantially, with the frequency of some of the codes doubling and even tripling (Table 1).

To better understand how this coding is done in the Police, we have also conducted several interviews with lead investigators responsible for the coding of the decisions. The interviews

[1] Married couples, couples living together, couples that have been married, have lived together or are having children together.

Decision	Official statistics	Estimated true value	Change (%)
Other	135 410	70 254	-48%
Crime cannot be proven	25 449	51 955	104%
Action is not a crime	10 613	34 139	222%
Suspect under 15	13 272	14 956	13%
Statute of limitation expired	933	1 030	10%
Minor offence	239	332	39%
Suspect dead	228	318	39%

Table 1. Decisions before and after recoding

revealed that there was no coherent view on how to utilize the codes, explaining the large variance in the utilization of the *Other* code at the different counties (Figure 1). As an example, in the Jönköping county all the interviewed persons said that the *Other* code is often preferred due to its flexibility. In contrast, the investigators in the Uppsala county thought that this code should be used sparsely. Differences also existed in the interpretation of the other available codes.

The main explanation for the variations in the usage of the decision codes is simple: There is no central guidance on how to use them. The police officers need to infer the correct usage from the names of the codes alone. For some codes, alternative sources of information than the Central Police Authority are available. However, this information is often unsystematic and incomplete.

Conclusions and discussion

We have assessed the quality of two code groups that constitute the building blocks of the official crime statistics in Swedish: the crime codes and the decision codes. For both groups, we have found a systematic over-usage of some of the general codes, leading to an uncertainty regarding the correctness of the statistics at the detailed level. Furthermore, a majority of the errors for both code groups were found whenever there were no clear administrative and/or judicial purposes for the codes.

An interesting finding regarding the coding of crime codes was that our results were similar to the ones reported in a previous study conducted by Statistics Sweden in 1978. The over-utilization of the *Other* codes, the misunderstanding of relationship status between victim and perpetrator in cases of assault, and the overall misclassification ratio for all reported offences were similar in both studies. Since then, additional instructions were included, computerized registration has been implemented, the Police was reorganized, and a whole new generation of police officers is in place. With this in mind, the quality issues associated with the crime codes seem to be systematic.

Regarding the decision codes, different opinions and practices at a local level on how to use the codes made this type of statistics nearly unusable at a national level. For the same reason, it is not possible to compare the work of police authorities in different counties. In the best of the cases, someone holding in-depth knowledge of the practices of a police authority in a specific county could perhaps analyze the development over time for certain decision types.

At a closer examination of the misclassified codes, we found that a large proportion of the errors occurred where there was no clear understanding of the administrative and juridical use

of the coding. Taking fraud as an example, no matter if a case is registered as C*omputer fraud*, *Fraud with use of false invoice* or *Other fraud*, the person responsible for the crime will still be convicted by the same paragraphs. A similar situation was observed for the decision codes, where the Police needs to give a motivation as to why they ceased a certain investigation. For many of the registering officers, the code *Other* with the possibility of phrasing a description in a free text, is an attractive option to do this. Thus, in this case, the need to statistically follow different motivations over time is competing with the need to be flexible and nuanced in explaining why a certain case is closed.

Although our study is focused on criminal data, the results are likely to hold relevance for other registry-based statistics as well. Considering the errors identified in our analyses, and their causes, we have formulated a few recommendations applicable in the broader context of statistics based on administrative sources:

- When creating, or describing statistics based on administrative sources, it is necessary to thoroughly analyze the processes that generate the raw data, using for example targeted quality studies.
- Quality issues are likely to be found where the administrative and/or juridical need is vague or non-existing. The statistical authority needs to map and estimate the impact of these errors, and actively inform users about this.
- General options, such as *Other type of fraud* or *Other type of robbery* are likely to be over-used. Whenever those options exist in a coding schedule, they need to be examined, so that the statistical authority can understand their impact on the final statistical product.
- The statistical authority needs to maintain close collaborations with the reporting agencies, in order to assure that instructions are clear and definitions well adapted to both the statistical needs and the administrative reality.

To conclude, we have shown a few examples where systematic errors led to coding inaccuracies, which in turn translated to large flaws in all the downstream statistics. This has had negative effects for the understanding of the distribution and development of certain crime types, and potentially harmful consequences for the strategic planning in the judicial sector. From the larger perspective of all official statistics, this clearly shows the relevance of analyzing the quality of the raw data, which is often neglected. In our view, only when getting a thorough understanding of the data formation processes, and the errors associated to this, will it be possible to fully exploit the wealth of data available for statistical production.

References

Bakker, F.M. Bart, 2012, *Estimating the validity of administrative variables*, Statistica Neerlandica Vol 66(1), p 8–17

Bakker, F.M. Bart and Daas, J.H. Piet, 2012, *Methodological challenges of register-based research*, Statistica Neerlandica Vol 66(1), p 2–7

Biemer, Paul and Lyberg, Lars, 2003, *Introduction to survey quality*. Wiley, Chichester, England

Groen A. Jeoffrey, 2012, *Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures*, Journal of Official Statistics, Vol 28(2), p 173–198

Holmberg, Anders, 2012, *Discussion on assessing quality of administrative data*. Statistica Neerlandica Vol 66(1), p 34-40

Statistics Sweden, SCB, 1978, *Fel vid kodning av brott – kvalitetskontroll av brottstatistiken*, Memo 1978:1, Stockholm

Statistics Sweden, SCB 1999, *Att mäta statistikens kvalitet,* Report 1999:3, Stockholm

Statistics Sweden, SCB, 2006, *Bakgrundsfakta, evalvering av utbildningsregistret*, Report 2006:4, Örebro

Statistics Sweden, SCB, 2007, *Slutrapport kodningsprocessen yrkesregistret. – Lotta P2*

SOU 2012:83, Final report from the investigation regarding statistics, 2012, *Vad är officiell statistik? En översyn av statistiksystemet och SCB*, Stockholm, Fritze

The national board of Health and Welfare, 2010, *Dödsorsaksstatistik, historik, produktionsmetoder och tillförlitlighet*. Collected from the home page of the national board of Health and Welfare. Produced in April 2010, collected in October 2011

Wallgren, Anders and Wallgren, Britt, 2007, *Register-based Statistics: Administrative Data for Statistical Purposes*, Wiley, Chichester, England

Zhang, Li-Chun, 2012, *Topics of statistical theory for register-based statistics and data integration*, Statistica Neerlandica Vol 66(1), p 41–63

Appendix

Estimating the net error rate in the stratified sample for *Other* codes

N_1 is the population for *Other* codes, as published in the official statistics.
N_2 is the population for the codes of all the defined crime types, as published in the official statistics.
$n_1$is the sample of O*ther* codes taken from the strata N_1.
$n_2$is the sample of codes for defined crime types taken from the strata N_2.
t_A= The amount of falsely assigned cases.

t_B= The amount of falsely exempted cases originally coded as a crime type from the population N_2.
The net error rate, $p_1 = \frac{t_A - t_B}{N_1}$
Falsely assigned offences:
$\hat{t}_A = \frac{N_1}{n_1}\sum_{i=1}^{n_1} A_i$ where $A_i = \begin{cases} 1\ if\ case\ is\ falsely\ assigned\ as\ type\ A \\ 0\ otherwise \end{cases}$

Falsely exempted offences:
$\hat{t}_B = \frac{N_2}{n_2}\sum_{i=1}^{n_2} B_i$ where $B_i = \begin{cases} 1\ if\ case\ is\ falsely\ exempted\ from\ type\ A \\ 0\ otherwise \end{cases}$

Estimating the net error rate in the stratified sample for defined codes

t_C= The amount of falsely exempted cases originally coded as a crime type from the population N_1.
t_D= The amount of cases from the population N_2 which was originally coded as the crime type for which the net error is being estimated.
The net error rate = $\frac{t_A - t_B - t_C}{t_D}$
Falsely assigned offences:
$\hat{t}_A = \frac{N_2}{n_2}\sum_{i=1}^{n_2} A_i$ where $A_i = \begin{cases} 1\ if\ case\ is\ falsely\ assigned\ as\ type\ A \\ 0\ otherwise \end{cases}$

Falsely exempted offences from the strata N_2:
$\hat{t}_B = \frac{N_2}{n_2}\sum_{i=1}^{n_2} B_i$ where $B_i = \begin{cases} 1\ if\ case\ is\ falsely\ exempted\ from\ type\ A \\ 0\ otherwise \end{cases}$

Falsely exempted offences from the strata N_1:
$\hat{t}_C = \frac{N_1}{n_1}\sum_{i=1}^{n_1} C_i$ where $C_i = \begin{cases} 1\ if\ case\ is\ falsely\ exempted\ from\ type\ A \\ 0\ otherwise \end{cases}$

The amount of offences originally coded as the type for which the net error is being estimated:
$t_D =$ The amount of cases originally coded as the crime type for which the net error is being estimated for.

Significance test for the net error rate

Confidence interval for rates in stratified samples:
The net error rate = Pi

$\frac{1}{N^2}\Sigma N_i^2 \frac{Pi(1-Pi)}{ni} = Var\ (P)$
Standard error= $\sqrt{Var(P)}$
Confidence interval =Net error $\pm$avarage error* Z
Z= tabular value according to normal distribution.

Test for change of distribution of offences before and after the coding

The test is based on the chi square properties. It has been calculated as follows:
$\chi^2_{(k-1)} = \frac{(O_i - E_i)^2}{E_i}$ The χ^2value has k-1 degrees of freedom and k represents the number of groups or categories in every test. O is the observed value, in this case the number of offences according to the correct coding, and E is expected value, in this case the number of offences according to the original coding done by the police.

If the χ^2 value is higher than the tabular value, then the difference in distribution before and after the coding is considered significant.