

EVALUATING THE REASONABLENESS OF DATA: USING CROSS RATIOS AND CHI-SQUARE MEASURES

Yoshio Akiyama¹, Samuel Berhanu²

ABSTRACT

The Uniform Crime Reporting (UCR) Program has devoted a considerable amount of resources in a continuous effort to improve the quality of data. The authors introduce and discuss the use of the Cross Ratios and Chi-Square measures to evaluate the reasonableness of data. UCR data is used to empirically illustrate the use of this approach.

KEY WORDS: Data Reasonableness; Data Quality; Cross Ratios, Chi-Square; UCR Program.

1. INTRODUCTION

There are a variety of reasons for bad quality of data in the UCR context. Institutionally, data could be misclassified or misrepresented, or the data collection process may have flaws. Unreasonable data may arise also from poor recording (both on paper and electronic forms). Any statistical program would suffer a loss of public confidence as a result of releasing unreliable data. Both the economic and social costs of correcting this problems are enormous. The UCR Program, being the major crime data source in the nation, is not immune from this type of a problem.

The UCR Program is founded upon the voluntary participation of the nation's law enforcement agencies. The quality, integrity, and accuracy of the data mostly rest upon the efforts made by each individual agency. Therefore, the program's capability and role are limited to indirectly detecting and rectifying errors or aberrations in reported data. For this purpose, a large portion of the UCR Programs resource is dedicated to data quality control.

1.1 Objectives

The main objective of this paper is to present statistical methodology developed to evaluate the reasonableness of UCR data. The specific objectives are to:

1. Introduce the use of Cross Ratios and Chi-Square tests to evaluate an agency's data.
2. Compute the test parameters using jurisdictional UCR data.
3. Discuss the implications in the context of data reasonableness review.

¹Senior Statistician, Federal Bureau of Investigation, J. Edgar Hoover Building, Room 11194, 935 Pennsylvania Avenue, Washington, D.C., USA 20535.

²Statistician, Crime Analysis, Research, and Development Unit, Criminal Justice Information Services Division, Federal Bureau of Investigation, 1000 Custer Hollow Road, Clarksburg, WV, USA 25306.

1.2 Data Source

The cross-sectional UCR murder data for the year 2000 are used in this study. Agencies with populations of 250,000 and over are considered for illustration. Three variables (the numbers of murders, arrests, and clearances) are used in computing the test statistics.

2. THEORETICAL BACKGROUND AND ESTIMATION PROCEDURE

2.1 Introduction

In the UCR Program, occasions have arisen that require tests of proportionality. There are a variety of reasons and scenarios where the need to employ some kind of statistical measure has become a necessity. These situations that necessitate a test of proportionality have been discussed in the data quality review of agencies' reports (Akiyama, Y., 2001). The following are some of the situations that allow close examination of the data.

1. There are agencies whose crime totals are in line with similar agencies. However, when the distribution of the data is examined, it could be substantially different.
2. An agency may have an annual total that indicates no conspicuous differences from similar agencies. However, the report may show excessively high monthly fluctuation beyond the normal seasonality.

2.2 Cross Ratios

Consider a $2 \times k$ table with its classes indexed by $i = 1, 2, \dots, k$. Let n_1, n_2, \dots, n_k ($n = \sum_{j=1}^k n_j$) be an agency numbers of observations associated to the classes and $p_i = n_i/n$ their proportions. Let π_i be the standard benchmark or expected proportion (arising from the similar agencies) for the i^{th} class and $e_i = n\pi_i$, $i = 1, 2, \dots, k$, denotes expected frequency for the i^{th} category. It is assumed that numbers in Table 1 are non-zero. The above notations give rise to the following two equivalent tables: one for the frequencies and the other for the proportions.

Table 1. The Frequency Distributions

	1	2	. . .	k	Total
Observed	n_1	n_2	. . .	n_k	n
Expected	e_1	e_2	. . .	e_k	n

Table 2. The Probability Distributions

	1	2	. . .	k	Total
Observed	p_1	p_2	. . .	p_k	1
Expected	π_1	π_2	. . .	π_k	1

(The number of observations = n)

The *cross ratio* from the i^{th} class to the j^{th} class is defined by

$$\theta_{ij} = \frac{p_i \pi_j}{p_j \pi_i} = \frac{n_i e_j}{n_j e_i} \quad [1]$$

Equation [1] shows that θ_{ij} is a *relative* measure and that

$$\theta_{ij} > 0, \theta_{ii} = 1, \theta_{ih}\theta_{hj} = \theta_{ij}, \text{ and } \theta_{ij} = \theta_{ji}^{&1}. \quad [2]$$

$\theta_{ij} = 1$ means that the classes i and j have the same proportions for the observed (the data to be tested) and the expected (the benchmark distribution).

2.3 Transformed Cross Ratios

The last identity $\theta_{ij} = \theta_{ji}^{&1}$ in the equation [2] shows that the cross ratios are *directional* in that the ratio “from i to j ” is the reciprocal of the ratio “from j to i .” However, the cross ratios should not be directional, since the transposition of i^{th} and j^{th} classes in Tables 1 and 2 does not affect the intrinsic proportionality of the two distributions. One of the ways to introduce a direction-free measure from the cross ratio θ_{ij} is to transform it to ϕ_{ij} as below:

$$\phi_{ij} = T(\theta_{ij}) = \frac{(\theta_{ij} + 1)^2}{\theta_{ij}} = \frac{(p_i \pi_j + p_j \pi_i)^2}{(p_i \pi_i)(p_j \pi_j)}. \quad [3]$$

The resulting ϕ_{ij} are called the *transformed cross ratios* with the following properties:

- $\phi_{ij} \geq 0$.
- ϕ_{ij} are direction-free, i.e., $\phi_{ij} = \phi_{ji}$, because $T(\theta) = T(\theta^{&1})$. Therefore, with ϕ_{ij} the proportionality is “between i and j ” instead of directionally “from i to j ”.
- $\phi_{ij} = 1$ if and only if $\theta_{ij} = 0$.
- The further the cross ratio θ_{ij} is away from 1, the further ϕ_{ij} is distanced from 0, i.e., we have
 - (a) $1 > \theta_{ij} > \theta_{hm} > 0$ implies $\phi_{hm} > \phi_{ij} > 0$,
 - (b) $\phi_{ij} = 0$ is the lowest value, and
 - (c) $\theta_{ij} > \theta_{hm} > 1$ means $\phi_{ij} > \phi_{hm} > 0$.

Further, we can extend cross ratios θ_{ij} as follows:

$$\text{Then, we have } \theta_{ihj} = \frac{\theta_{ih}}{\theta_{hj}}. \quad [4]$$

$$\theta_{ijj} = \theta_{ij}, \theta_{iij} = \theta_{ji} = \theta_{ij}^{&1}. \quad [5]$$

and

$$T(\theta_{ihj}) = \frac{(\theta_{ih} + \theta_{hj})^2}{\theta_{ij}} = \frac{(p_i p_j \pi_h^2 + \pi_i \pi_j p_h^2)^2}{(p_i p_j \pi_i^2)(\pi_i \pi_j p_h^2)}. \quad [6]$$

We define the *generalized transformed cross ratios* as:

$$\phi_{ihj} = T(\theta_{ihj}). \quad [7]$$

The following relationship holds:

$$\phi_{ij} = \phi_{ih} \phi_{hj} \geq \phi_{ih} \geq \phi_{hj} \text{ \& } \phi_{ihj}, \text{ for any } i, h, \text{ and } j. \quad [8]$$

Compare [8] with the simpler relationship $\theta_{ij} = \theta_{ih} \theta_{hj}$.

Finally, we have the following relationship:

$$\frac{p_j}{\pi_j} = \prod_{i=1}^k p_i \theta_{ji}. \quad [9]$$

2.4 The Relationship of Chi-Square Test to Cross Ratios

For Tables 1 and 2, the chi-square test of fit is given by:

$$X^2 = \sum_{i=1}^k \frac{(n_i - e_i)^2}{e_i} = n \left[\sum_{i=1}^k \frac{(p_i - \pi_i)^2}{\pi_i} \right]. \quad [10]$$

Let

$$A = \sum_{i=1}^k \frac{(p_i - \pi_i)^2}{\pi_i}, \quad [11]$$

so that $X^2 = nA$. X^2 is known to be distributed according to the χ^2 -distribution with $(k-1)$ degrees of freedom. We have the following identity:

$$A = \sum_{k \leq j > i \leq 1} p_i p_j \varphi_{ij} = \sum_{k \leq j > i \leq 1} p_i p_j T(\theta_{ij}). \quad [12]$$

Therefore, we obtain the following relationship between X^2 and the transformed cross ratios:

$$X^2 = n \left[\sum_{j>i} p_i p_j \varphi_{ij} \right]. \quad [13]$$

Below, the meaning of the product $p_i p_j$ contained in X^2 is considered. Two independent multinomial experiments with the parameter (p_1, p_2, \dots, p_k) are given by

$$1 = (p_1 + p_2 + \dots + p_k)^2 = \sum_{i=1}^k p_i^2 + 2 \left(\sum_{j>i} p_i p_j \right).$$

Therefore, the product $p_i p_j$ ($i \dots j$) is viewed as the observed *relative weight* of the classes i and j . Then, X^2 is a function of the sample size n , the observed relative weights $p_i p_j$, and the transformed cross ratios φ_{ij} . X^2 has $k(k-1)/2$ terms in [13].

2.5 Choosing a Test of Proportionality

The two measures, cross ratios and chi-square measure, are fundamentally different statistics. In this section, we compare the two and point out the differences. First, the cross ratios are relative measurements, while X^2 is an absolute measurement: X^2 takes the observation size, n , in consideration while cross ratios do not. As a result, cross ratios fail to distinguish two situations where proportions/distributions are the same but have different n -values.

Secondly, X^2 reflects the observed relative weights $p_i p_j$, but the cross ratios do not depend on them. For example, illustrating through the case $k = 2$, the following two tables (Tables 3 and 4) have the same cross ratios but different values for $p_1 p_2$.

Table 3.

Classes	1	2	Total
Observed	30 ($p_1 = 1/4$)	90 ($p_2 = 3/4$)	120
Expected	48 ($\pi_1 = 2/5$)	72 ($\pi_2 = 3/5$)	120

Table 4.

Classes	1	2	Total
Observed	24 ($p_1 = 1/5$)	96 ($p_2 = 4/5$)	120
Expected	40 ($\pi_1 = 1/3$)	80 ($\pi_2 = 2/3$)	120

Both tables have the same cross ratios $\theta_{12} = 0.5$ and the same transformed cross ratios $\phi_{12} = 0.5$. However, their relative weights are different: Table 3 has $p_1 p_2 = 3/16$ while Table 4 has $p_1 p_2 = 4/25$. By (13), the two tables have different X^2 s: $X^2 = (120)(3/16)(0.5) = 11.25$ for Table 3 and $X^2 = (120)(4/25)(0.5) = 9.60$ for Table 4. Therefore, in this example, the cross ratios do not distinguish the two tables, but X^2 does.

Thirdly, the above two examples might have given an impression that X^2 is a more sensitive measure than the cross ratios. However, we can generate an example where the cross ratios distinguish the given two tables but X^2 fails to do so. This is shown below.

Table 5.

Classes	1	2	Total
Observed	30 ($p_1 = 1/4$)	90 ($p_2 = 3/4$)	120
Expected	24 ($\pi_1 = 1/5$)	96 ($\pi_2 = 4/5$)	120

Table 6.

Classes	1	2	Total
Observed	5 ($p_1 = 1/4$)	15 ($p_2 = 3/4$)	20
Expected	8 ($\pi_1 = 2/5$)	12 ($\pi_2 = 3/5$)	20

$\theta_{12} = 4/3$ and $\phi_{12} = 1/12$ in Table 5, while $\theta_{12} = \phi_{12} = 1/2$ in Table 6. Since cross ratios (and transformed cross ratios) are different, they serve to distinguish the two tables. However, the chi-square test does not distinguish the two tables since both have the same X^2 . Fourthly, X^2 incorporates all cross ratios into one expression through [13]. However, cross ratios are separate measures.

As mentioned earlier, X^2 is the product of the number of observations n and the relative indicator A . In order to address the dominance of a large n , the following discussions are made. Let

$\Delta = \text{Max} \{ \theta_{ij} \}$. Then, $\theta_{ij} \leq \frac{1}{\Delta}$, because $\theta_{ij} = \theta_{ji}^{&1} \leq \Delta^{&1}$. Therefore,

$\Delta \leq \theta_{ij} \leq \frac{1}{\Delta}$ for all i and j . In terms of the transformed cross ratios, this implies that

$$0 \leq \phi_{ij} \leq \frac{(\Delta + 1)^2}{\Delta}, \quad [14]$$

because $T(\Delta) = T(\Delta^{&1})$. Therefore, if $\Delta \neq 2$, for example, we have

$$A = \sum_{j>i} p_i p_j \phi_{ij} \leq \frac{(\Delta + 1)^2}{\Delta} \sum_{j>i} p_i p_j = \frac{(\Delta + 1)^2}{\Delta} \frac{1}{2} \sum_{i=1}^2 p_i^2 < \frac{(\Delta + 1)^2}{2\Delta} \leq \frac{1}{4}, \quad [15]$$

$0 \leq A \leq 1/4$, if $\Delta \neq 2$.

Equation [15] suggests the general range of variation for A and limits its range. Equation [13] implies, on the other hand, that even a very small A (that indicates a high proportional conformity to the expected distribution) might be negated by a large n , i.e., a small A can be magnified by an overriding magnitude of n . Therefore, in situations where n is very large, a large X^2 results from a large n . When n is large, we consider the alternative measure $X^2 = \sqrt{n} A$. Outliers consist of agencies that have large values in X^2 .

2.6 Transformations of X^2

As the degree of freedom k increases, the square-root and cube-root transformations of X^2 yields normal distributions. First, the square root $\xi = \sqrt{2X^2}$ is approximately normally distributed with mean $\sqrt{2k+1}$ and unit variance (Stuart, A. and Ord, J. K, 1987). In notation,

$$z_1 = \frac{\sqrt{2X^2} - \sqrt{2k+1}}{\sqrt{2}} \sim N(0, 1). \quad [16]$$

Secondly, the cube root $\eta = \sqrt[3]{X^2/k}$ is approximately normally distributed with mean $1 + 2/(9k)$ and variance $2/(9k)$ expressed as:

$$z_2 = \frac{\sqrt[3]{\frac{X^2}{k}} - \left(1 + \frac{2}{9k}\right)}{\sqrt{\frac{2}{9k}}} \sim N(0, 1) \quad [17]$$

Since normal variables are more tangible, reports from agencies can be compared in terms of z_1 or z_2 (instead of X^2) for the same result. When k is not large, both normal approximations may be inaccurate. However, they are usable for the UCR purpose, since neither z_1 nor z_2 changes agency ranking or relative position established by X^2 . The outlier test will be “one-tailed” in that agencies having high z_1 or z_2 are selected.

3. COMPUTATIONS AND RESULTS

3.1 Computation Procedure

Based on the mathematical and statistical algorithm discussed in Section 2, software has been developed to calculate the statistics, θ , ϕ , z , and χ^2 . The program reads data for the number of murders ($i = 1$), arrests ($i = 2$), and clearances ($i = 3$). The software compares the proportionality of each agency against the total ratio's for the whole group (67 agencies combined). Estimates of θ , ϕ , and χ^2 are sorted and printed for each variable (murder, arrest, and clearance).

3.2 Results and Discussion

Out of 67 agencies 10 agencies are chosen for the sake of illustration and discussion. Tables 7 through 9 describe cross ratios (θ_{ij}), transformed cross ratios (ϕ_{ij}), chi-square (X^2), and standardized values (z). The first table gives rise to the second table, which in turn is used to compute the test statistics in the third table. Agencies with excessively high chi-square (or standardized) values (Table 9) are designated as outliers in terms of the proportionality and distribution test. Population size does not have impact on the value of test statistic, because the issue at hand is the proportionality of the three variables (numbers of offenses, clearances, and arrests). In Table 9, the value $X^2 = 1,662.82$ for Agency 10 is excessively high, while the values for the other nine agencies are closer to each other. Therefore, data from Agency 10 is considered an outlier requiring further examination. The extremely high values of X^2 are interpreted as a sign for large deviations from the other agencies pattern of proportionality in the numbers of crimes, clearances, and arrests. The above approach does not directly utilize cross ratios as test statistics. However, they are useful (a) in the initial detection of data anomaly and (b) in the identification of which variables are too high or low. This is because cross ratios are the building blocks of computing the chi-square statistic as earlier shown. The observation (b) is of particular importance: Cross ratios are used in as diagnostic means since the test statistic does not indicate which variable is causing large increase in the values of X^2 .

Table 7. Cross Ratios

Agency	Population	Offense (θ_{12})	Clearance (θ_{23})	Arrest (θ_{13})
1	888,632	0.47195	1.44821	0.6835
2	1,020,055	0.67252	1.29237	0.8691
3	676,701	0.83310	1.80433	1.5033
4	1,266,132	0.74031	2.06233	1.5268
5	1,193,440	0.85865	1.82973	1.5711
6	263,937	0.43669	1.35787	0.5930
7	883,621	0.36050	1.19413	0.4305
8	422,266	0.57408	1.54746	0.8884
9	1,119,580	0.69203	5.42234	3.7524
10	972,390	0.91838	0.17312	0.1590

Table 8. Transformed Cross Ratios

Agency	Population	Offense(ϕ_{12})	Clearance(ϕ_{23})	Arrest(ϕ_{13})
1	888,632	0.59082	0.13872	0.1466
2	1,020,055	0.15946	0.06614	0.0193
3	676,701	0.32976	0.51282	0.0183
4	1,266,132	0.09019	0.54722	0.1817
5	1,193,440	0.02327	0.77626	0.2076
6	263,937	0.72664	0.09432	0.2794
7	883,621	1.13440	0.03156	0.7534
8	422,266	0.31601	0.19368	0.0140
9	1,119,580	0.13705	3.60677	2.0189
10	972,390	0.00725	3.94959	4.4489

Table 9. Chi-Square and Standardized Values

Agency	Population	χ^2	Z
1	888,632	5.09	1.4581
2	1,020,055	5.31	1.5256
3	676,701	5.72	1.6494
4	1,266,132	6.92	1.9878
5	1,193,440	7.39	2.1113
6	263,937	7.77	2.2089
7	883,621	13.50	3.4642
8	422,266	18.20	4.3009
9	1,119,580	105.81	12.8150
10	972,390	1662.82	55.9363

3.3 Conclusion

In this paper we have attempted to show how the cross ratios, transformed cross ratios, and chi-square statistics relate to each other and can be used in detecting outliers in data proportionality. The chi-square measure is used as an index without emphasizing the corresponding probabilities. In terms of probabilities, both $Prob(X^2 \leq 105.81)$ for Agency 9 and $Prob(X^2 \leq 1662.82)$ for Agency 10 are practically zero and are indistinguishable. This measure is very useful in determining the quality of data.

Since the UCR Program is a voluntary program that does not lend itself to direct agency data auditing, the above indirect reasonableness checks based on standardized approach is deemed very useful. For the maximum effectiveness of this method, it is important that “similar” agencies are defined using refined criteria including the level of urbanization, geography, agency types, and demographic compositions.

REFERENCES

Akiyama, Yoshio (2001), “UCR Data Quality Control,” unpublished report, Department of Justice, Federal Bureau of Investigation, Washington, D.C.

Stuart, A. and Ord, J. K. (1987), *Kendall's Advanced Theory of Statistics*, Vol. 1, Oxford University Press.