Methods of Data Quality Control:
For Uniform Crime Reporting Programs

Cross-Sectional Outlier Detection
Longitudinal Outlier Detection
Proportional Outlier Detection

Yoshio Akiyama
Sharon K. Propheter

Criminal Justice Information Services Division
Federal Bureau of Investigation

April 2005

Table of Contents

Executive Summary

Since the Uniform Crime Reporting (UCR) Program is founded upon the voluntary participation of the Nation's law enforcement agencies, the integrity and the accuracy of its data rest upon two factors: (a) the efforts of individual law enforcement agencies in reporting accurate data and (b) the Program's capability for data quality control (detecting and rectifying errors or aberrations in reported data).

Data quality control consists of the following dimensions:

- (Prior to Reporting) Errors are reduced through training and education.
- (At Data Processing) At the data processing stage, UCR conducts logical edits and statistical reasonableness tests.
- (At Data Aggregation) UCR reviews the soundness of statistical output.
- (After Data Processing) Reported data are examined through the Quality Assurance Review (QAR) process.

Let us consider the data quality procedures at the data processing stage. The *logical edits* relate to the elimination of incorrect data at the data processing stage. Software and human reviews are combined to assess the validity of reported data in terms of logic. The *reasonableness reviews* are a statistical data checking in the aggregate as opposed to logical errors. Outliers, rather than errors, are detected here. This note describes the algorithms for data reasonableness reviews useful in detecting outliers.

Algorithms consist of the following three components:

- Cross-Sectional Outlier Detection.
- Longitudinal Outlier Detection.
- Proportional Outlier Detection.

Outliers are selected by comparing an agency with similar agencies. This is achieved by dividing agencies into strata based on population size, urbanization, agency type (sheriff or city police), and geographical location. In Part 3, the scheme of stratification is described as a basis of outlier detection and subsequent data estimation. Each agency belongs to a unique stratum.

Cross-sectional outlier detection is described in Part 4, wherein agencies' crime rates (for a prescribed month or a year) are compared to the *median crime rate* of the stratum used as the *norm*. The algorithm used for this purpose is the ratio of the agency's crime rate to the median crime rate of a stratum. The philosophy behind this algorithm is that (even though an agency's

crime rate can differ from similar agencies) an *excessive* difference from the median indicates potential data anomaly.

Longitudinal outlier detection examines an individual agency's self-consistency over time in reported crimes. It compares the preceding year's number with the current year's. Since excessive or unlikely fluctuations for consecutive years suggest data aberrations or changes in reporting practices, agencies showing too large increases or decreases in reported crimes are identified as outliers. While the longitudinal algorithm measures "jumps in time" pertaining to a single agency, agency jumps are compared within the stratum.

The third component of the outlier detection algorithms is the *proportionality* test. Proportionality includes such examples as the balance of weapon distributions, monthly distributions of reported crimes, proportions of simple and aggravated assaults, distributions of the amount of monetary losses, proportions of violent and property crimes, and the proportion of offenses and clearances. The proportionality test measures the deviations of agencies' distributions from the stratum distribution taken as the norm.

The designation of an agency's data as an outlier does not mean that it is invalid or faulty. Since algorithms are mechanical, the uniqueness of individual agencies chosen as outliers must be further examined by humans. Therefore, the final decisions on individual agencies are made by human scrutiny, i.e., outliers will become usable or unusable after the manual examinations.

The results and methods of data quality reviews cannot be separated from the subsequent data estimations. Therefore, conscious efforts have been made to tie the proposed quality reviews closely to the data imputation algorithms [Reference 10]. See Section 7 and Appendix E.

Part 1. The Scope of Data Quality Control

The UCR Program is founded upon the voluntary participation of the Nation's law enforcement agencies in reporting jurisdictional crime incidence. Since there is no (mandatory) data auditing program to directly validate individual agencies' data and reporting practices, the integrity and the accuracy of data rest upon the efforts of individual law enforcement agencies in reporting accurate data and the Program's capability of data quality control (e.g., detecting and rectifying errors or aberrations in reported data).

The UCR data quality has many aspects starting from training and education prior to data submission, followed by components related to the reliability, accuracy, and logic of data at the processing stage. The UCR data quality control has many components related to the processing of the incoming reports, data utilization, and presentation. Data processing involves examining the reliability, accuracy, and logic of data. Additional components review policies, methodologies, utilization, and presentation. QARs are conducted to further the education and training aspect of data quality control. All phases are important since data quality relates to raw data as well as statistical output. The object of this report is to address the aspect related to quality control of the reported data at the stage of processing.

Training and education are conducted in concert with state UCR Programs through ongoing local training and educational sessions and through communications that address UCR standards, guidelines, policies on classifications and scoring rules, etc. The training and education are *proactive* data quality control measures and focus on the reduction of common reporting errors.

QAR staff visit volunteer agencies, review their reported data and reporting practices, and prepare statistical reviews relating to the accuracy of their data. These reviews can be accumulated to provide global adjustments to the published statistics. This assessment for the already published statistics illustrates the *retrospective* dimension of the QAR Program.

The specific data editing process and the data reasonableness review fall between the above two programs (training/education and QAR) and relate to the verification of reported data at the data processing stage. The data editing process reviews reported data for possible *logical errors* and *inconsistencies*. For the National Incident-Based Reporting System (NIBRS), it is in this process that error messages and flags are attached to reported incidents. For example, error messages and warning flags are associated with inadmissible data combinations (e.g., monetary loss in assault), incompatible data (e.g., the date of arrest preceding the date of incident), missing mandatory data (e.g., not reporting the agency's Originating Agency Identifier number or weapon usage for aggravated assault).

In the NIBRS, the data editing addresses the internal accuracy and consistency of

"individual" incidents rather than the "aggregated" statistical data. The NIBRS data editing process is documented in the NIBRS Volume 4*: Error Message Manual* (December 3, 1999). There is also computer software that edits UCR Summary reports.

There are a number of differences between the specific data editing and the data reasonableness review. The following are some of their differences:

- **Logical Correction vs. Plausibility Review:** The data editing process relates to *logical errors*, while the reasonableness review relates to the *plausibility* of reported data. For example, a NIBRS report that associates monetary loss to an assault incident contains a *logical* error since this incident should have been reported as a robbery incident. This type of a problem is addressed in the data editing process. On the other hand, if an agency reports 100 murders but only 5 larcenies in a given month, the reasonableness of the relative frequencies of the two offense categories is questioned. The latter example constitutes a data plausibility issue and, therefore, is addressed in the data reasonableness review.
- **Individual Data vs. Aggregate Data:** In NIBRS, the data editing process is oriented to individual/specific data elements within a given incident. In the UCR Summary system, the data editing process relates to the compatibility and consistency of individual numbers within and among Summary reporting forms (Return A form, Supplements to Return A, etc.). On the other hand, the data reasonableness review addresses data *in aggregate* (i.e., statistics). Therefore, the reasonableness review of NIBRS data is based on statistics rather than on an individual incident.

As mentioned earlier, this paper describes methods of a data reasonableness review or the algorithms used in detecting outliers. This is achieved in Parts 3–6. In Part 3, the scheme of stratification is described as a basis of outlier detection. Each agency belongs to a unique stratum based on population size, urbanization, agency type (sheriff or city police), and geographic location. The stratum defines similar agencies to which an agency is compared. Central values (e.g., average and median) for the stratum are used as the *norms* in evaluating agencies' reports. It is an algorithm assumption that the majority of agencies send reasonable reports. Although norms are influenced by reports that may later be designated as outliers or unusable, they constitute valid gauges in evaluating the individual agencies' reports.

In Part 4, the method for detecting cross-sectional outliers is described. For each crime category, agencies' crime rates are cross-sectionally compared to the *median crime rate* of the stratum. In Part 5, the algorithm used to examine the agency's self-consistency is explained. For a given agency, a crime category, and a time period, the method compares the preceding year's report with the current year's report. Since excessive fluctuations in the number of crimes for consecutive years suggest possible data aberration or changes in reporting standards, those

agencies showing large "jumps" in the number of crime incidents are chosen as outliers. The measures of fluctuations (jumps) for individual agencies are compared and ranked within the stratum. This test is longitudinal in that the measurement of jumps arises from the agency's historical data alone. However, it is partly cross-sectional since the measured jumps are compared with other agencies.

The third component of the outlier detection algorithms described in Part 6 pertains to *proportionality*. By the proportionality test, we evaluate the deviations of agencies' data distributions to the stratum distribution (norm). The following are examples of proportionality issues (for each agency):

- Weapon distribution for a violent crime.
- Monthly variations of reported crimes.
- Ratio in the numbers of simple and aggravated assaults.
- Distribution of monetary losses.
- Ratio of violent and property crimes.

In proportionality testing, the reference of comparison (the norm) is the average distribution for the stratum. Distributional outlier detection may overlap with the cross-sectional and longitudinal tests. For example, if an agency's January report shows an inordinately high number of robberies in the monthly variation, the cross-sectional outlier algorithm for January data would also detect this anomaly.

Part 7 addresses the procedures needed in implementing these algorithms. Examples of issues discussed are:

- How often should the outlier algorithms be run (e.g., monthly, quarterly, semiannually, and cumulatively)?
- When should the final human decisions on data acceptability be made (e.g., at the end of a calendar year, before publication, or at each outlier review)?
- How should unaccepted outliers be handled? For example, should unaccepted outliers be included in establishing the stratum norm in subsequent outlier detections?
- What kinds of status indicators should be used to record the human decisions reached on outliers?
- How should late reports submitted after outlier tests be handled?

Chapter 8 describes data reviews for large agencies and, finally, Part 9 describes the method of arrest data review.

Part 2. Methodological Remarks

Issues that should be considered in developing outlier tests are described below:

- The first issue involves establishing a measurement to determine how much a given datum is away from the established norm, i.e., the extent of deviation from the norm. The choice of such measurement is a primary concern since an inappropriate measure could select less serious deviations over more serious ones or select marginal outliers over genuine outliers.

- The second issue in developing outlier tests is the introduction of a *cut-point* (the amount of admissible deviation). A cut-point is a threshold point for a report to become an outlier. For example, if the upper cut-point of a measurement is c, then an agency's report that measures beyond c will be an outlier. In choosing a cut-point, there is no objective criterion as to how much deviation is too much, i.e., there is no universally accepted criterion as to what constitutes an excessive and unreasonable deviation from the norm. Therefore, the choice of a cut-point is an empirical or managerial matter. In data reasonableness review, threshold points (cut-points) are associated to each measurement. Based on the measurement and cut-point, outliers are selected representing the offense category pertaining to an agency's report for a given period of time. For example, an outlier could be the number of robberies reported by an agency for the month of January.

- The third issue is determining what to do with the selected outliers. Since there is no hands-off quality control, the next step of data reasonableness review is a manual/human examination of the outliers' acceptability. This decision relies upon known factors and conditions that might have influenced the agency's crime reports. Human decisions incorporate such factors as agencies' responses to inquiry letters (i.e., the agency's account of data anomaly), communications with state UCR staff, the agency's historical patterns of crime, special conditions that prevailed in the agency during the period under review, changes made in the agency's record-keeping and reporting systems, etc.

- It is important to note that outliers selected by the algorithms do not automatically become unacceptable data. A large deviation from the norm does not necessarily imply that reported data are faulty or unusable. Outliers are nothing but "candidates" that should be classified as acceptable or unacceptable. After human examinations, outliers could prove to be acceptable. Outliers become unaccepted only when they are judged unreasonable and not well-founded. The above observation indicates that the status of being an outlier is temporary and the designation of data acceptability is an ultimate decision.

The requirements for the automated outlier detection process are as follows:

- **Streamlining the data quality review process:** By automatically screening out reasonable reports as nonoutliers, algorithms streamline the data quality review process. *False alarms* (selecting usable reports as outliers) present no serious problems because human reviews would establish their usability.

- **Not overlooking questionable reports:** The algorithms should not miss any report that shows unlikely, unreasonable, or excessive deviations from the norm. Algorithms should pick all questionable reports. *False negatives* (overlooking unacceptable data) are not allowed.

The outlier algorithm must present two quantities for the selected outliers:

(a) The extent of numerical deviation from the norm.
(b) The level of seriousness (or rarity) of such a deviation from the norm.

The need for (a) and (b) is demonstrated through agencies' crime rates x_i and the median m for the stratum. The ratio $y_i = x_i / m$ indicates how much the agency's crime rate is deviated from the median rate and, therefore, satisfies the first requirement (a). However, this measurement y_i does not satisfy (b) since there is no information on the distribution of y_i. The "rank of y_i" satisfies (b) but not (a). The measurement y_i and the rank of y_i jointly express (a) and (b).

It is possible that a distribution function $f(y)$ can be fitted to the data y_i. Then, the probability $P_f(y > y_\alpha) = \alpha$ represents the significance or the rarity of the deviation y_i. Even when the function $f(y)$ faithfully represents the data y_i, agencies selected on the basis of the significance $P_f(y > y_\alpha)$ do not differ from those selected by the rank of y_i. When the fit of $f(y)$ is loose, the proportion of agencies selected by the criterion $y_i > y_\alpha$ does not reflect the significance α. The choice of the type of distributions is not automatic. Therefore, the seriousness (or rarity) of the observation y_i as measured through the distribution $f(y)$ is not accurate.

Information (a) is indispensable. Without (a), an agency's report could be selected as an outlier based on a minor deviation from the median. For example, consider the case where observations are clustered around the median. When most y_i's are closely clustered around one, an agency's report may become an outlier due to a minor deviation from the others. In such a situation, however, the agency's report should not be an outlier. Depending solely on the *significance* of the deviation (whether it is measured by $f(y)$ or by the rank) is not a sound approach without being accompanied by (a). In the above example, the ratio y_i and its rank give

both sets of information, (a) and (b), since y_i indicates the *measure of deviation* from the median and the rank tells *how rare* such a deviation is.

Agencies are expected to deviate from the established norms. This is natural and no agency should become an outlier for deviating from the norms. Individual agencies are unique and expected to be different from the norm. This is illustrated below by using year-to-year changes in the level of crime. What we are looking for in outliers is not whether an agency's report indicates a statistically significant difference from the norm. Significant shifts in the crime level routinely happen for agencies. For a shift to be statistically significant, two standard deviations would be sufficient. But, for a year-to-year crime shift to be called unlikely, excessive, or questionable, a larger deviation (higher than two standard deviations) would be needed.

To further illustrate the above point, let us consider a change from 100 burglaries to 115 burglaries. Although this increase could be statistically significant, this in itself may not warrant the designation of an outlier and, therefore, we must further examine the agency's report. As mentioned above, the procedure is not about *significant* changes but about *excessive* changes. A shift from 100 burglaries to 300 would be judged excessive and chosen as an outlier. The latter report represents a 200-percent increase in burglary, which is considered unlikely under a normal condition and would require further scrutiny.

The Methodological Epilogue (Part 10) provides additional technical explanations.

Part 3. Stratifying into Similar Agencies

In detecting outliers, the algorithms compare a given agency's report to those of similar agencies. The choice of similar agencies is made by a stratum to which the agency belongs. A stratification is therefore a scheme of partitioning agencies into similar agencies.

In developing a stratification, it was decided that a scheme should incorporate the following four factors:

- The jurisdictional size (i.e., the agency's population size).
- The geographic area where an agency is located (i.e., geographic regions/U.S.).
- The degree of urbanization (i.e., in Metropolitan Statistical Areas or not).
- The agency type (i.e., sheriff's office vs. police department).

These four variables have been historically used as a basis of UCR crime estimation and are considered to represent the minimum set of requirements. The new scheme, differing only by geographical designation (state vs. region/U.S.), is based on stable and available factors that do not require constant updates or assessment of their applicability.

The stratification scheme is based on data external to the UCR Program but not based on agencies' reported crime data. If reported crime or arrest data were used as one of the factors in stratification, the resulting stratification would partially be reduced to a classification through self-declaration. By design, each stratum contains an adequate number of agencies. In this way, norms used in outlier detections will have sufficient meaning and stability, and all strata have large enough numbers of reports to define outliers.

The following stratification is used as the basis for defining similar agencies in the outlier detections and lead to subsequent data estimation. There are 33 strata defined in Table 3.A.

Table 3.A. New Stratification

Strata Number	MSA*	Population** (in thousands)	City/County	Region***
1	Y	$250	City	U.S.
2	Y	100 to 250	City	U.S.
3	Y	50 to 100	City	U.S.
4	Y	25 to 50	City	1
5	Y	25 to 50	City	2
6	Y	25 to 50	City	3
7	Y	25 to 50	City	4
8	Y	10 to 25	City	1
9	Y	10 to 25	City	2
10	Y	10 to 25	City	3
11	Y	10 to 25	City	4
12	Y	2.5 to 10	City	1
13	Y	2.5 to 10	City	2
14	Y	2.5 to 10	City	3
15	Y	2.5 to 10	City	4
16	N	$ 10	City	1
17	N	$ 10	City	2
18	N	$ 10	City	3
19	N	$ 10	City	4
20	N	2.5 to 10	City	1
21	N	2.5 to 10	City	2
22	N	2.5 to 10	City	3
23	N	2.5 to 10	City	4
24	Y	n/a	County	1

25	Y	n/a	County	2
26	Y	n/a	County	3
27	Y	n/a	County	4
28	N	n/a	County	1
29	N	n/a	County	2
30	N	n/a	County	3
31	N	n/a	County	4
32	n/a	< 2.5(non-zero)	City	U.S.
33	n/a	0	n/a	U.S.

* Y = Yes, N = No.

** "*x* to *y*" means "greater or equal to 1,000 *x* but less than 1,000 *y*."

*** **Northeastern States (Region 1):** Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont, New Jersey, New York, and Pennsylvania.
Midwestern States (Region 2): Illinois, Indiana, Michigan, Ohio, Wisconsin, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota.
Southern States (Region 3): Delaware, District of Columbia, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, West Virginia, Alabama, Kentucky, Mississippi, Tennessee, Arkansas, Louisiana, Oklahoma, and Texas.
Western States (Region 4): Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming, Alaska, California, Hawaii, Oregon, and Washington.

Stratum 33 (the stratum for zero-population agencies) will be excluded from cross-sectional outlier checks because crime rates are not computable for agencies with zero populations. Also, this stratum is not amenable to other outlier tests because it contains many jurisdictions with diversified missions and territories.

Part 4. Method for Detecting Cross-Sectional Outliers

Three types of algorithms will be discussed in Parts 4 through 6. A cross-sectional comparison is a major cornerstone of outlier algorithms. It should be noted that the cross-sectional component requires automation and is not feasible in the manual approach.

Although agencies belonging to the same stratum are considered similar, they have different jurisdictional populations. For a given stratum (for example, Stratum 20, Table 3.A.), populations for big agencies are four times the smaller agencies. Therefore, comparing reported *volumes of crimes* would result in the selection of false outliers because small agencies with small volumes and large agencies with large volumes tend to be picked as outliers. Since outlier detection should not be a reflection of jurisdictional size, the cross-sectional comparison is more reasonably based on the jurisdictional crime rate since it is more standardized by design with respect to the population.

For a given stratum of similar agencies and for a specific crime category, let n be the number of agencies whose reports are cross-sectionally compared, and let x_i, $i = 1, 2, \ldots, n$, be the crime rate for the i-th agency. Let m be the *median crime rate* of the stratum. We consider the following n ratios:

$$y_i = \frac{x_i}{m}, \quad i = 1, 2, \ldots, n. \tag{4.1}$$

The ratio $y_i = 1.8$, for example, means that the crime rate of the agency is 1.8 times greater than the median. The cross-sectional outlier algorithm bases the outlier designation on the ratio y_i and its rank $r(y_i)$. The agencies that showed highest and lowest ranks (for example, the top and the bottom one percent) are chosen as outliers.

Note: When the number of agencies is small, one percent of n may not select an agency. Therefore, it is stipulated that $0.01\,n$ is raised to a whole number (i.e., rounded up). Also, agency ratios too close to the median (i.e., $1 - \varepsilon < y_i < 1 + \varepsilon$ for a prescribed small ε) should not be selected as an outlier even when it is included in one percent of the selected agencies. Agencies are designated as outliers if they rank very high (or low) and, at the same time, their y_i are outside the band $1 - \varepsilon < y_i < 1 + \varepsilon$.

Below, the ratio y_i is explained as the *cross-product ratio*. Let C_i be the number of crimes and P_i be the population of the i-th agency. Let C_{med} and P_{med} be the number of crimes and the population for the median agency if n is odd. For an even n, C_{med} and P_{med} are the average of the middle two. Then, we have the following table.

Table 4.A.

	Number of Crimes	Population
The *i*-th Agency	C_i	P_i
Median agency	C_{med}	P_{med}

It is noted that

$$y_i = \frac{x_i}{m} = \frac{C_i \, P_{med}}{P_i \, C_{med}}. \tag{4.2}$$

Therefore, y_i is the *cross-product ratio* for Table 4.A. In this table, the per-person crime rates C_i / P_i and C_{med} / P_{med} are the *odds*. We have $y_i = 1$ if and only if the two rows in the tables have the same proportions.

As mentioned before, whether the status of an outlier changes to the status of unacceptable data is individually determined by the staff. Based on such human decisions, the computer will print the following information relating to outliers selected for human examination:

- The agency's name that reported the outlier.
- The test month.
- The year of the data.
- The stratum number to which the agency belongs.
- The offense category for which the outlier designation was made.
- The value of x_i.
- The median value *m*.
- The value of the ratio y_i.
- The rank $r(y_i)$.
- The number *n* of agencies compared.
- The type of algorithm that made the outlier designation (i.e., the cross-sectional test in this case).
- Other information to be later designated.
- Data Source Indicator: NIBRS or Summary data.
- The graphical representation of the distribution of $\{y_i\}$, *i=1, 2, ..., n*.

Note: The choice of the median *m* could be replaced by such factors as the crime rate of the stratum or the average of the agencies' crime rates. Although they are usable, it was considered that the median is less susceptible to extreme or unreasonable crime rates x_i.

Part 5. Method for Detecting Longitudinal Outliers

In this section, the method for detecting longitudinal outliers is discussed. The algorithm is designed to detect excessive changes in the agency's reports during the most recent two periods. This algorithm is viewed as a test for an agency's self-consistency.

For a given agency and data category, let X denote the number of offenses for the current year and Y for the preceding year. X and Y are regarded as variables. It is assumed that the report $Y=Y_0$ from the agency was accepted last year (as being a reasonable report). Therefore, the longitudinal self-consistency of the current report $X=X_0$ (i.e., the consistency and the compatibility of the new report $X=X_0$ with the accepted report $Y=Y_0$ for the preceding year) is under review.

The variables X and Y are assumed to be Poisson-distributed, and $X=X_0$ and $Y=Y_0$ are regarded as the observed values. In notation, we write:

(5.1) $$X \sim Po(\lambda), \quad Y \sim Po(\mu),$$

where λ and μ are Poisson parameters. To test the null hypothesis $H_0 : \lambda = \mu$, it is known that the conditional distribution of X given $T = X + Y$ is binomially distributed with the parameter $p = 1/2$:

(5.2) $$P(X \geq X_0) = \frac{1}{2^T} \sum_{X=X_0}^{T} \binom{T}{X}, \text{ where } X_0 = 0, 1, \ldots T.$$

See [Reference 4: pp.140-143].

We choose the condition $T = X_0 + Y_0$. Since T is generally a large number, the above distribution (5.2) can be approximated by the standard normal distribution $N(0, 1)$ as:

(5.3) $$P(X \geq X_0) \approx N(z \geq z_0),$$

where

$$z = \frac{X - pT}{\sqrt{Tp(1-p)}} = \frac{2X - T}{\sqrt{T}} = \frac{2X - (X_0 + Y_0)}{\sqrt{X_0 + Y_0}}.$$

Note: When T is small, the longitudinal test has a limited meaning.

Under the condition $T = X_0 + Y_0$, the test value used against the null hypothesis $H_0 : \lambda = \mu$ is

$$z_0 = \frac{X_0 - Y_0}{\sqrt{X_0 + Y_0}} = \frac{X_0 - Y_0}{\sqrt{T}}. \tag{5.4}$$

When the value z_0 is too large or too small, we reject the null hypothesis. The report $X=X_0$ is considered inconsistent with the agency's historically accepted data $Y=Y_0$. See Appendix A.

The denominator of (5.4) could be zero. However, the following modification avoids the situation of the vanishing denominator:

$$z_0 = \frac{X_0 - Y_0}{\sqrt{X_0 + Y_0 + 1}}. \tag{5.5}$$

Formula (5.5) will be used as the measure of deviation. As explained in Part 4, there remains a problem of defining what constitutes an excessive change. There is no accepted theory or logic that determines how many standard deviations (away from the norm) make a change excessive or unreasonable. For this purpose, the algorithm ranks the agencies' measurements within the stratum and chooses unlikely deviations (e.g., the top and the bottom one percent). An outlier thus chosen shows not only a significant difference but also an excessive deviation from the preceding year's report.

Formula (5.5) should be compared with the percent change indicator. The latter is a totally relative indicator and ignores the volume of crimes involved in the computation. For example, the change in the number of burglaries from 100 (last year) to 130 (current year) represents a 30-percent increase. Likewise, the change in the number of burglaries from 450 (last year) to 585 (current year) represents a 30-percent increase. Therefore, the percent change indicator does not distinguish the two scenarios. Based on (5.5), however, these two scenarios provide totally different values:

$$z_0 = \frac{130 - 100}{\sqrt{130 + 100 + 1}} = \frac{30}{\sqrt{231}} = 1.974 \text{ (for the first scenario).}$$

$$z_0 = \frac{585 - 450}{\sqrt{585 + 450 + 1}} = \frac{135}{\sqrt{1{,}036}} = 4.194 \text{ (for the second scenario).}$$

Formula (5.5) not only distinguishes scenarios (that have the same percent changes) but also assigns higher z_0 values to larger volumes. See [Reference 1: Part I] for more details.

Note: The standard normal large-sample test is based on the continuity-corrected statistic

$$z = \frac{X - Y}{\sqrt{X + Y}} \pm \frac{1}{\sqrt{X + Y}}$$

where $+1/\sqrt{X + Y}$ is used if $X < Y$, while $-1/\sqrt{X + Y}$ is used if $X \geq Y$. See [Reference 8: Page 227].

Output for the outliers from the longitudinal test are similar to those mentioned in Part 4 for the cross-sectional outliers. To be more precise, it includes the following:

- The agency's name that reported the outlier.
- The test month.
- The year of the data.
- The stratum number to which the agency belongs.
- The offense category for which the outlier designation was made.
- The value of z_0.
- The rank $r(z_0)$.
- The number n of agencies compared.
- The type of algorithm that made the outlier designation (i.e., the longitudinal test in this case).
- Other information to be later designated.
- Data Source Indicator: NIBRS or Summary data.
- The graphical representation of the distribution of the z_0's.

Part 6. Method for Detecting Distributional Outliers

It is possible that the total reported numbers are in concert with other similar agencies, but their breakdowns (e.g., the distribution of weapons used, the monthly breakdown in the number of crimes, the proportion of simple assault to aggravated assault, the proportion of violent crimes vs. property crimes, and the proportion of crimes, clearances, and arrests) are substantially different from the norms established through the stratum of similar agencies. A proportional/distributional outlier means a report with an excessive deviation from the norm in terms of its breakdown or frequency distribution. This deviation in the reported data proportionality can occur in a number of ways as explained in Part 1. Two examples follow:

Example 1: The number of robberies for an agency is in line with those for similar agencies. However, the breakdown of weapon usage is substantially different from others.

Example 2: An agency has reported data for 12 months, and the annual total indicates no conspicuous difference from similar agencies. However, high monthly fluctuations cannot be explained by seasonality or by normal monthly variations.

We consider the following two tables. In Tables 6.A. and 6.B., π_i represents expected proportions and $e_i \text{ ' } \pi_i n$ the expected numbers, where n is the total number of observations for the agency.

Table 6.A.
The Frequency Distributions

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

Table 6.B.
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)

In the above tables, the term "observed" denotes a distribution to be tested (an agency's report) and "expected" means a benchmark distribution (i.e., the norm established by the stratum).

The first problem is to assess which observed proportion (distribution) is more deviated from the expected proportion (distribution). For Tables 6.A. and 6.B., the chi-square test of fit is given by

$$X^2 = \sum_{j=1}^{k} \frac{(n_j - e_j)^2}{e_j} = n \left[\sum_{j=1}^{k} \frac{(p_j - \pi_j)^2}{\pi_j} \right]. \tag{6.1}$$

Let

$$A = \sum_{j=1}^{k} \frac{(p_j - \pi_j)^2}{\pi_j}, \tag{6.2}$$

so that $X^2 = n A$. X^2 is distributed according to the χ^2-distribution with (k-1) degrees of freedom and is used for the proportionality test.

X^2 is a product of the volume factor n and the relative factor A, i.e., X^2 takes into consideration the observation size n and the relative factor A that arises strictly from proportions (A is unrelated to the volume of observations). Although the factor A occupies a higher weight in X^2 when n is relatively small, X^2 is strongly influenced by n when n is a large number.

When n is very large (as in the case of larceny), the formula (6.1) can be modified to

$$\hat{X}^2 = \sqrt{n}\, A \tag{6.3}$$

to reduce the impact of n while retaining the impact of the relative factor A. It is equivalent to considering a smaller observation size $\sqrt{n}$ instead of the actual (and larger) observation size n. In the distributional outlier test, therefore, $\hat{X}^2$ is the measurement used. In application, the values $\hat{X}^2$ are computed for agencies and ranked (or shown in percentiles) in selecting distributional outliers.

Note: See [Reference 1] for the relationships between A and the cross-product ratios.

Note: It is known that (as the degree of freedom k increases) the square root and cube root transformations of X^2 yield normal distributions. First, the square root $\sqrt{2X^2}$ is approximately normally distributed with mean $\sqrt{2k - 1}$ and unit variance (Fisher). See [Reference 5, Page 508]. In notation,

$$z_1 = \sqrt{2X^2} - \sqrt{2k - 1} \;\dot{\sim}\; N(0, 1). \tag{6.4}$$

Secondly, the cube root $\sqrt[3]{X^2/k}$ is also approximately normally distributed with mean $1 - 2/(9k)$ and variance $2/(9k)$ (Wilson and Hilferty - 1930). See [Reference 5, Page 509].

In notation,

$$(6.5) \qquad z_2 \;\dot{=}\; \frac{\sqrt[3]{\frac{X^2}{k}} - \left(1 - \frac{2}{9k}\right)}{\sqrt{\frac{2}{9k}}} \;\dot{\sim}\; N(0,1).$$

See [Reference 5: pp. 508-513]. Since normal variables are more common, reports from agencies can be compared in terms of z_1 or z_2 (instead of $\hat{X}^2$). When *k* is not large, both normal approximations are inaccurate.

The second problem is to decide how much deviation is unreasonable. As before, this is done by ranking the values $\hat{X}_i^2$, i=1, 2, ..., n, and selecting, e.g., the top and the bottom one percent of the agencies.

As in Parts 4 and 5, the computer will print for the staff similar information as before once outliers are selected for human examination. See pages 13 and 16 for details.

As mentioned in Part 1, a report selected as an outlier due to an excessive deviation in proportionality could in certain situations be selected through a cross-sectional outlier or a longitudinal test. To illustrate, if the proportion of aggravated assault is too low in comparison with the number of simple assaults via a proportionality test, the cross-sectional test or the longitudinal test would detect the low number of aggravated assaults (without referencing simple assault). However, the proportionality test is considered effective in testing the reasonableness of crime subcategories (e.g., the weapon usage in a robbery).

Part 7. Implementation of Statistical Data Quality Control

The object of Part 7 is to describe how data quality control algorithms are tied to the UCR operations in data quality reviews. As mentioned earlier (Part 2), the statistical data reasonableness review goes through the following two steps:

First Step (Outlier Detection Algorithms): The first component is the automated selection of outliers via algorithms. The outliers pertain to the agency's prescribed data for a given time frame. For example, an outlier could be the number of robberies reported by an agency for a prescribed month. A report becomes an outlier when the deviation is too large when compared to the experiences of similar agencies. Let *X* represent a report for a given agency, data category, and time frame. If *X* is not selected as an outlier, *X* is assigned a code N (= Non-Outlier). The report *X* with N is not subjected to human scrutinies. If *X* is selected as an outlier, it is given the code O (= Outlier) and moves to the second step.

Second Step (Data Acceptability Review by Humans): When the code O (= Outlier) is associated to *X*, human reviews and examinations of data begin. This is a manual stage of quality assurance in terms of data acceptability into the database. This decision relies upon known factors that might have influenced the agency's data. Human decisions incorporate such factors as an agency's responses to inquiry letters (i.e., the agency's account of data anomaly), communications with state UCR staff, the agency's historical patterns of crime, special conditions that prevailed on the agency during the period, changes made in the agency's record-keeping system or its reporting practices, etc.

It is important to note in the above discussions that outlier *X* selected by the algorithm does not automatically become an incorrect or bad report. A large deviation from the norm does not necessarily imply that reported data are faulty or not usable. Outliers become unacceptable or not usable only when they are judged (by the staff) as being too excessive, unreasonable, and not well-founded to accept it as a part of the UCR data.

Although additional reviews can be conducted, the following three steps of data quality reviews are performed at a minimum:

A. Initial Monthly Reviews: When the staff consider that a sufficient number of agencies have submitted data for a given month, an initial monthly review takes place. Outliers are chosen from the reports that are available at the time, while some agencies' reports may still be unavailable at this juncture. The major thrust of the initial monthly review is cross-sectional. For crimes that are infrequently committed or for smaller agencies, the monthly review may be inconclusive or sometimes meaningless due to the small numbers involved. Human examinations can take care of such individualities.

Since the initial monthly tests are intermediary in nature and conducted without the whole set of reports for the year, the decisions made at the initial monthly reviews may require changes. However, these preparatory reviews are necessary to make the final decisions well-informed and meaningful. Only the agencies that report data monthly (as opposed to nonstandard submissions such as annual or quarterly submissions) are involved in the initial monthly reviews. Agencies practicing nonstandard submissions receive only annual reviews.

Again, let *X* be a report from an agency for a given month and data category. If *X* is not an outlier, the code N (= Non-Outlier) is associated with *X*, and the report is accepted. If *X* is an outlier, it is designated as O (= Outlier), and the second step (review by humans) starts. At the second step, human decisions (or indecisions) are recorded using the following *status indicator codes*:

A = the outlier is accepted for full use.
S = the outlier is accepted for self-representation.
U = the outlier is unacceptable.
D = the decision is deferred.

If the status code A (= Acceptable) is associated with the monthly report *X*, the report (although it is an outlier for the month and for the data category) is considered acceptable for full use. These data are used for all releases, compilations, and publications (assuming that other monthly reports from the agency are acceptable). It will be seen that reports with A are used for imputing other agencies' missing monthly data for the prescribed data category.

The status indicator code S (= Self-Representation) is a variation of A in conjunction with data estimation. It denotes that *X* is acceptable and released (with footnote) in all output. However, in the data estimation, a report with S should represent only the submitting agency and should not be used for imputing other agencies' missing or rejected data. Typically, S is given in special situations where extreme data (outliers) are caused by certain known conditions (such as riots, bombings, etc.) that the jurisdiction experienced during a particular reporting period. Although reports with S are valid data, they are considered too atypical to impute or estimate missing agencies' data. Therefore, the difference between A and S is that, in the process of data estimation, S is only for self-representation while A constitutes a basis for imputing missing or rejected data.

An agency's outlier report *X* is given the status indicator code U (= Unacceptable) when the departure from the norm is too excessive and cannot be supported by reasonable explanations. At the close of the reporting year, monthly data designated U will be reevaluated along with other monthly data, and a final decision will be made regarding the acceptability of the agency's annual reports. Agencies' reports that have the status code U for any month will be

reexamined irrespective of the reasonableness of the annual total and other monthly reports. See Appendix B for the flowchart of the ramifications of the initial monthly review.

The status code D (= Deferred) denotes that the decision was deferred or no decision was possible within the framework of available information. When no human decision is reported by the staff with respect to the outlier *X*, this code is given by default. D can be altered to other codes (A, S, or U) when more information is subsequently obtained and a definitive human decision is reached.

B. Final Monthly Review: At the close of the reporting year, another monthly review is conducted. This end-of-the-year monthly review is called the final monthly review. No human decisions are exercised regarding the outliers arising from this stage. All existing monthly reports are included in the final monthly review, i.e., they are used to compute the norms of comparison in the final monthly review. Of the total of such agencies' reports, outliers are chosen. The objects of the final monthly review are to:

a. Assess the reasonableness of the monthly reports that were submitted late and not included in the initial monthly review (i.e., missing reports submitted later).

b. Assess the reasonableness of the updated monthly reports. This includes the following situations:

- The report *X* was selected as an outlier at the initial monthly review.
- Human decision (A, S, D, or U) was made to *X*.
- *X* was updated after a human decision was made.
- No human decision was made to the updated report.

Decisions reached prior to the final monthly review (i.e., N, A, S, and U) are not altered at the final monthly review unless they were updated after such decisions.

c. Reevaluate the outliers that were D at the initial monthly review and where no updating has been made.

d. Evaluate updated monthly reports.

- The report X was N (= Non-Outlier) at the initial review.
- X was updated subsequently by a new report.

At the final monthly review, the following designations are made:

- For (a): Reports that were M (= Missing) at the initial monthly review but were submitted later are examined. The code MO (= Missing-Outlier) is given if the

reports submitted late are outliers. Otherwise, the code N (= Non-Outlier) is given.

- For (b): Updated monthly outliers mentioned in (b) are examined. The code DO (= Deferred-Outlier) is given if the report is still an outlier. The code N (= Non-Outlier) is given if the report is not an outlier.

- For (c): Outliers with D (but with no subsequent updating) are examined. An outlier detection test will assign N or DO to the report.

- For (d): NO (= N became O) is assigned if the updated report is an outlier. Otherwise N (=Non-Outlier) is given.

At the final monthly review, numerical values are associated with the final outcomes. They represent the relative seriousness (in terms of data reasonableness) of the monthly reports. The values are: N = A = 0, S = 1, MO = NO = 2, DO = 3, and U = 6. Although the numerical values are subjective, they are deemed to reflect the relative seriousness of the respective situations. For example, DO = 3 is higher than MO = NO = 2 because the former was designated twice as an outlier, while the latter is a one-time outlier. No numerical value is given to M. Appendix C describes the flowchart of the possible ramifications of the final monthly review.

C. Annual Review: As mentioned previously, numerical values are assigned to represent the gravity, or the severity, of the problems in terms of data reasonableness for a given monthly report. Outlier tests and subsequent human scrutinies can be conducted at any time (not just at the initial and the final monthly reviews), and the results of such review can be used to modify human decisions made earlier. However, the final decisions regarding the acceptability and the usability of the agencies' annual reports are made at the annual review process. At the annual review, all three algorithms (cross-sectional, longitudinal, and proportional tests) may be used.

1. **Incomplete Crime Categories:** For a given data category, if an agency does not have 12 monthly reports at the end of the year, the data category is called incomplete. Incomplete data categories are automatically coded as UP (= Unpublishable Crime Category). If an agency has an unpublishable data category (i.e., classified as UP), the agency's reports are coded UR (= Unreleasable Agency).

   UR (= Unreleasable Agency) means that the agency is excluded from agency listings. In UCR, agency listings are statistical tables that present individual agencies' data by their jurisdictional names. Table 8 (Offenses Known to the Police) in the annual publication, *Crimes in the United States*, is an example of an agency listing. An agency with any data category classified as UP becomes UR (= Unreleasable Agency) and is

excluded from agency listings. This is based on the philosophy that actual data for individual jurisdictions should be distinguished from estimated data.

2. **When the Annual Total is a Non-Outlier:** For a given data category, let us assume that the agencies under consideration have reported all 12 months (i.e., no missing months), and *Y* is the agency's annual total. As already mentioned, each of the 12 monthly reports have one of the numerical values: 0, 1, 2, 3, or 6.

   If the annual total *Y* is N (= Non-Outlier), the sum of the monthly status codes is examined. Agencies whose sum of numerical values is 6 or above will undergo further human examinations.

   Annual totals (that are not outliers) that received 6 points or above are individually reviewed by the staff to determine their reasonableness. If for a given data category, the agency's annual total *Y* and the monthly reports (that received the total of 6 points or above) are judged publishable for a given crime category, the code P (= Publishable Crime Category) is assigned. On the other hand, if the agency's crime total *Y* (not an outlier) and 12 monthly reports (that received 6 or above) are considered unpublishable for a given data category, the code UP (= Unpublishable Crime Category) will be assigned. It should be noted that P and UP are codes given to each data category.

   If an agency that does not submit reports monthly (e.g., submitting quarterly, semiannually, or annually) receives the non-outlier designation for the reported annual total *Y*, the agency's data are considered as P.

3. **When the Annual Total is an Outlier:** When the agency's annual report *Y* is an outlier, it is automatically reviewed by the staff irrespective of the agency's modes of data submission (i.e., a monthly, quarterly, semiannual, or annual total submission). The final decision regarding the usability of the agency's annual total for the year is made in conjunction with the annual total *Y* (which is an outlier) and the records of monthly reasonableness reviews. The possible ramifications of the annual review are described in Appendix D.

What has been discussed so far pertains to a given category of crimes. In order for an agency's reports to be releasable for the year, all Part I crimes must receive P. However, this has an exception of UPP (= Unpublishable Crime Category due to Policy), the designation of a crime category (such as forcible rape data for Illinois and Delaware) that are declared *a priori* unpublishable due to the policy or system's discrepancy. If all Part I crimes are judged P or UPP, the agency is called R (= Releasable Agency), and its data will be included in agency listings. Otherwise, the agency receives UR (= Unreleasable Agency), which happens when UP

is assigned to one or more crime categories. Reports from agencies that are classified as UR are excluded from the individual agency listings.

Note 1: If an agency with R requests not to be individually released, it is excluded from the agency listings.

Note 2: If the status indicator code S is given to any month for a releasable agency, an appropriate footnote will accompany the agency's data.

4. **Imputation of Monthly Data:** Data quality reviews are tied to subsequent data estimation. At the close of the reporting year, the final monthly outlier run is made. Based on this final run, monthly data imputations are made on some of the monthly reports. The following monthly reports receive data estimations:

- All reports with missing months.
- All monthly reports designated as U (= Unacceptable) when it is a part of UP.
- All monthly reports designated as UPP (= Unpublishable Crime Categories due to Policy). Forcible rape data for Illinois and Delaware are examples.
- No monthly data belonging to P are imputed (even when the report is from unreleasable agencies).
- All monthly reports classified as DO (= Deferred-Outlier), MO (= Missing-Outlier) or NO (= N became O) are estimated if they belong to UP whose annual totals are O (= Outlier) or I (= Incomplete).
- For agencies that submit data in a nonstandard manner (e.g., quarterly, semi-annual, or annual submissions), monthly estimations are made if the crime category is UP or UPP, but no monthly estimations are made for P.

See Appendix E for the summary of the rules for monthly crime estimations. (Further details of this topic will be addressed in the forthcoming paper, "A Method of Data Estimation.")

The quality control methods (outlier detections) addressed so far are limited to cross-sectional, longitudinal, and proportional outlier tests. As these methods become operational, there will be more dimensions added to the total data quality assurance process. Although the routine data quality review as described in this paper is limited to the number of monthly and annual data, a more refined special review process can extend this scope and be executed toward larger agencies (e.g., agencies with over 100,000 inhabitants).

The outlier tests are founded upon the philosophy that the majority of the agencies within the stratum submit reasonable reports, so the norms established through the stratum provide

reasonable yardsticks to assess the statistical reasonableness of individual agencies' reports. However, this assumption sometimes must be confirmed as data analyses and data quality reviews reveal scenarios that remind us of the need for such confirmations. For example, the *UCR State Program Bulletin* (November 30, 1998) states that NIBRS permits the reporting of both larceny and burglary in the same incident if they are separate offenses. In most cases, however, burglary involves the crime of larceny. If an agency improperly reports both a burglary and a larceny, the larceny count becomes inflated. Therefore, an agency should not report both the larceny and the burglary unless they are separate offenses. Depending on the prevalence of the improper reporting of burglary and larceny, the norm could become a distorted benchmark. Strictly speaking, a complete data review should consist of the following three dimensions: logical error checks, statistical reasonableness reviews, and adjustments for anecdotal instances of distorted benchmarks.

Part 8. Data Reviews for Large Agencies

Each of the three components of outlier detection is based on a single variable. The cross-sectional test is based on the agency's crime rate divided by the median crime rate. See (4.1). The longitudinal test is based on the *z*-value computed from the agency's numbers of crimes for consecutive years. See (5.4) and (5.5). The proportionality test is based on the modified chi-square measure. Irrespective of the power of these outlier detection algorithms, measuring the total complexity in the fluctuations of data is beyond a mechanical process (represented by outlier detection algorithms).

A small number of large agencies (jurisdictional populations over 50,000 regardless of NIBRS or Summary reporting) represent a large portion of the Nation's crimes. Therefore, applying more extensive methods of data quality control to large agencies is expected to enhance the data quality by:

- Extending the scope of applications for algorithms (e.g., increasing the check points for the distributional outlier test and extending the distributional test to a variety of situations),
- Increasing the categories of quality control models beyond the three components (cross-sectional, longitudinal, and distributional),
- Increasing the level of communications with large agencies,
- Acquiring more refined information for individual agencies,
- Securing monthly reports from large agencies to avoid missing data, etc.

As mentioned earlier, the distributional (proportional) outlier test can be used for multiple purposes. It can be used to check the distributions of data such as:

- Simple assaults and aggravated assaults.
- Violent crimes and property crimes.
- Exceptional clearances and clearances by arrest.
- Crimes completed and attempted.
- Recovered and unrecovered vehicles.
- The proportion of multiple vs. single-offense incidents.
- Monthly variations of the number of crimes.
- The proportion of crimes, clearances, and arrests for a given crime category.
- The distribution of offense locations.
- The distribution of the monetary losses in property crimes.
- The distribution of weapons used.
- The distribution of property losses.
- The distribution of victims by age, sex, and race.

- The distribution of injury types.
- The proportion of offenders, victims, and arrestees.
- The distribution of the victim-to-offender relationships.

The quality control models can be expanded to areas not covered by the three components of outlier tests. For example, the test can be applied to the time lapse patterns from theft to recovery and from crime incident to clearance, the average monetary losses in property crime, time series for crimes and arrests, and graphic displays of data. See for example [Reference 9].

Regular communications with large agencies and accumulation of information relating to each agency (such as its record-keeping system and special conditions prevailing within the agency) would refine the quality control process for large agencies. The expanded data quality control for large agencies has to be developed and applied with constant human involvement.

Part 9. The Review of Annual Arrest Data

In principle, arrest data should be submitted monthly. However, there are exceptions to this principle. A limited number of agencies report arrest data in a nonstandard manner. For example, an agency may submit only an annual arrest total. The review of annual arrest data relates to both types of reports.

Review of arrest data is conducted for annual arrest totals. At the end of the reporting year, arrest data from agencies reporting via NIBRS are converted to Summary statistics. In the UCR Summary system, the Age, Sex, Race, and Ethnic Origin of Persons Arrested (ASR) are collected for adults and juveniles, known as Adult ASR and Juvenile ASR. The Adult ASR form contains 27 crime categories for Part I and Part II crimes (26 crime categories if the drunkenness category is not a crime for the state). The Juvenile ASR form contains the maximum of 29 crime categories (or 28 if drunkenness is not a crime for the state).

The review of annual arrest data is conducted for individual crime categories. To be specific, the annual arrest data review for a given crime category is applied to agencies, except those in Stratum 33 (zero-population agencies), that satisfy the following conditions:

- Agencies submitted 12 monthly arrest reports for the crime category (or submitted arrest data on a nonstandard basis but covered the total year for the crime category, e.g., one annual report that included the combined arrest data for the year).
- Agencies submitted both Adult and Juvenile ASR forms for the crime category under review.
- Agencies' arrest reports included age and sex breakdowns for the crime category under consideration.

Reporting of the arrestee race is not required for the annual review. Also, it is not required that drug abuse violations and gambling contain arrest data breakdowns for subcategories.

Based on the stratification introduced earlier, agencies' arrest totals for a given crime category are compared with similar agencies within the stratum. The cross-sectional outlier test is applied as before. The agencies' arrest rates per 100,000 inhabitants are divided by the median arrest rate of the stratum, and the resulting quotients are compared in terms of their ranks as was done for the crime statistics. (See Part 4.) The longitudinal outlier test is applied to the volumes of arrests reported for the consecutive years (this year against the preceding year). The method of comparison is the same as in the crime reasonableness review. (See Part 5.) Finally, for Parts I and II crimes, the distributional outlier detection is applied. Outlier tests can be expanded to additional situations relating to NIBRS data.

Crime categories chosen as O (= Outlier) by any of the above outlier tests are examined by the staff and classified into one of the three categories: A (= Acceptable), S (= Self-Representation), and U (= Unacceptable). If the report is not an outlier, the code N (= Non-Outlier) will be assigned. For a given crime category, therefore, agencies are grouped into the following five classes:

- N = Non-Outliers.
- A = Outliers that are accepted for full use.
- S = Outliers that are accepted only for self-representation (in the arrest estimation process).
- U = Outliers that are unacceptable.
- E = Excluded from the annual arrest data review (including incomplete annual data, reports missing, etc.).

The above designations are tied to agencies and crime categories. See Appendix F for the ramifications of the annual arrest data review.

An agency's arrest report for a given crime category is "P (= Publishable Crime Category)" only if it belongs to N, A, or S. Otherwise, "UP (= Unpublishable Crime Category)" is assigned. It will be seen in the forthcoming paper [Reference 10] that the arrest data estimation for a given crime category is based on all agencies that reported arrest reports classified as N or A.

An agency with P for all crime categories is coded as R (= Releasable Agency). Otherwise, it receives UR (= Unreleasable Agency). When arrest statistics are released by individual agencies' names, only the releasable agencies are listed.

Part 10. Methodological Epilogue

This section explains the rationales used in developing the three outlier detection methods described in this paper: cross-sectional, longitudinal, and proportional. An attempt is made to discuss the underlying thought and intent behind these models, their limitations, and the misapplications associated with them.

For the following discussions, randomly choose a stratum from the first 32 strata. Although there are 33 strata in the stratification, the last stratum (Stratum 33 of zero-population agencies) is excluded from the scope. The agencies in a given stratum will be indexed by *i*, where $i = 1, 2, \ldots, n$. Let t_i be a *measurement* for the *i*-th agency for a given time frame. The following text reflects slight changes in the notation.

- **For Cross-Sectional Test:** $t_i = x_i / m$, where x_i is the crime rate of the agency and *m* is the median crime rate of the stratum. See (4.1).

- **For Longitudinal Test:** $t_i = (y_2 - y_1) / \sqrt{y_1 + y_2 + 1}$, where y_1 and y_2 are the numbers of reported crimes (for the *i*-th agency) for the preceding and current years. See (5.5).

- **For Distributional Test:** $t_i = \sqrt{n_i} \sum^{k} (p_{ij} - \pi_j)^2 / \pi_j$, where n_i is the number of observations, *k* is the number of categories, and p_j and π_j are actual and expected proportions. See Part 6.

In outlier reviews, the test was about the null hypothesis, $H_i : \theta_i \le \theta_0$ for a given agency *i*, where θ_0 is the norm of the stratum. H_i is rejected only when the significance is *excessively* low. It is noted that outlier detection is not a hypothesis testing in the regular sense which rejects the hypothesis at a prescribed level α. In fact, many agencies are expected to have parameters that are different from the stratum norm θ_0. Only when t_i is *excessively* large, the agency having the excessive t_i is declared an outlier.

The proposed outlier algorithms avoided the use of the cut-point. Instead, outliers are picked according to a fixed percent in the ranked data: $t_{(1)} \ge t_{(2)} \ge \ldots \ge t_{(n)}$. The use of the ranked data is a nonparametric approach and selects an intended number of agencies as outliers.

Cross-Sectional Test: In choosing a measurement for cross-sectional outlier detection, agency statistics such as the number of crimes, the crime rate, and the rank in the crime rate can be considered. The rank may serve partially as a basis of agency-to-agency comparisons but retains no information on the measurement itself. For example, an agency ranking of "1" could arise from a small deviation from the norm when all agencies cluster around the norm.

The number of crimes is not a comparable measurement for outlier detections since it ignores the jurisdictional size. If outlier selection is based on the number of crimes, larger or smaller agencies would become outliers.

The crime rate is a more standardized measurement, although it is not a perfect device for comparison. The crime rate was examined as the measurement for the cross-sectional test statistic and was considered to have a Poisson distribution. The measurement $t_i = x_i / m$ (where x_i is the agency crime rate and m is the median) was chosen as the test statistic.

Poisson postulates are as below:

- Crime incidents in nonoverlapping time intervals are independent.
- The probability of crime incidents in a time interval depends on the length of time.
- There is a parameter λ such that the probability of one crime incident in a time interval of length h is equal to $P_1(h) = \lambda h + o(h)$.
- The probability of two or more crime incidents in a time interval of length h is very small, i.e., $P_n(h) = o(h)$, $n = 2, 3, \ldots$.

Based on the above postulates, it was considered that the crime rate is crudely Poisson distributed, i.e., $X_i \sim Po(\lambda_i)$, where λ_i is a parameter in the third postulate.

When an agency becomes an outlier based on the ranking, $t_{(1)} \le t_{(2)} \le \ldots \le t_{(n)}$, an assessment can be made if the selection is legitimate based on the *value* of t_i (rather than on the *rank* of t_i).

Longitudinal Test: In selecting a measurement for longitudinal outlier detection, the following agency measurements were initially considered:

- The *percent change* in the number of crimes.
- The *difference* between the numbers reported for consecutive periods, etc.

The *percent change* is not a viable basis of comparison since a small number tends to fluctuate more easily than a large number. Therefore, if outlier detection is based on the percent change, smaller agencies in the stratum are more likely to become outliers than the larger agencies. On the other hand, the *difference* in the numbers of crimes between the consecutive periods depends on agency size; larger agencies tend to become outliers more frequently than smaller agencies. Measurements based on the volume are generally not viable for outlier detection since they are too strongly tied to the agency size.

Starting with the assumption that the number of crimes is Poisson distributed, the measurement $t_i = (x_2 - x_1) / \sqrt{x_1 + x_2 + 1}$ was chosen as the basis of longitudinal measurement. The conditional distribution of x_2, given the value of the sum $x_1 + x_2$, is

binomially distributed and, therefore, can be normally approximated.

After outliers are chosen by ranking t_1, t_2, . . . , t_n, the legitimacy of the choice of the outlier can be assessed by the knowledge of its distribution, which gauges how excessive and extreme the selected outlier is.

Distributional Test: The development of the distributional algorithm was similar to the above cases. It is evident that statistics such as

$$\chi_i^2 = \sum_{j=1}^{k} \frac{(n_{ij} - e_j)^2}{e_j} \quad \text{or} \quad A_i = \sum_{j=1}^{k} \frac{(p_{ij} - \pi_j)^2}{\pi_j} \tag{10.1}$$

are either too volume-dependent or too relative (i.e., ignore the size of observations).

The Pearson's chi-square statistic χ^2 is a product of the volume n of observations and the strictly relative quantity A arising from the proportionality of the *observed* proportions and the *expected* proportions, i.e., $\chi^2 = n A$. The quantity A offers a partial basis of comparison but ignores the underlying size n of observations. The product $\chi^2 = n A$ can be strongly influenced by a large n.

Therefore, the test statistic

$$t_i = \sqrt{n_i} A_i = \frac{1}{\sqrt{n_i}} \sum_{j=1}^{k} \frac{(n_{ij} - e_j)^2}{e_j} \tag{10.2}$$

was chosen as the measurement. The test statistic t_i (10.2) is viewed as a chi-square statistic based on the reduced $\sqrt{n_i}$ observations.

Appendix A
Difference of Poisson Variables

This appendix corroborates the choice of statistic for the longitudinal test. See (5.4) and (5.5).

Let $\xi = X - Y$ be the difference of the two Poisson variables X and Y; it is recalled that X is the reported number for the current year and Y is for the preceding year. ξ is not a Poisson variable since it can take negative values.

Assuming the independence of X and Y, ξ is known to have the following probability density function:

(A.1). $$f(\xi) = e^{-2a}\left[\sum_{i=0}^{4} \frac{a^{\xi+2i}}{(\xi+i)!\, i!}\right], \quad \xi = 0, \pm 1, \pm 2, \pm 3, \ldots,$$

where a is the common parameter in the null hypothesis $H_0 : \lambda = \mu (= a)$. See [Reference 5: Example 4.5, pp.126-127].

It is known that all odd cumulants of ξ are zero (in notation, $\kappa_{2r+1} = 0$) and even cumulants are $2a$ ($\kappa_{2r} = 2a$). See [Reference 5: Example 11.17, p. 366]. Using the relationships between the cumulants and the moments, we see that ξ has the mean $E(\xi) = \kappa_1 = 0$ and the variance $V(\xi) = \kappa_2 = 2a$. See [Reference 5: p. 87]. Therefore, the standardized form of ξ is

(A.2) $$z = \frac{\xi - E(\xi)}{\sqrt{V(\xi)}} = \frac{\xi}{\sqrt{2a}} = \frac{X - Y}{\sqrt{2a}}.$$

Replacing $2a = \lambda + \mu$ by the minimum variance bound (MVB) estimator $X_0 + Y_0$, we see the test value of (5.4) becomes

(A.3) $$z_0 = \frac{X_0 - Y_0}{\sqrt{X_0 + Y_0}}.$$

For the MVB estimator of a Poisson distribution, see [Reference 6: Example 17.8, p. 617].

Remark: Since X and Y are independent,

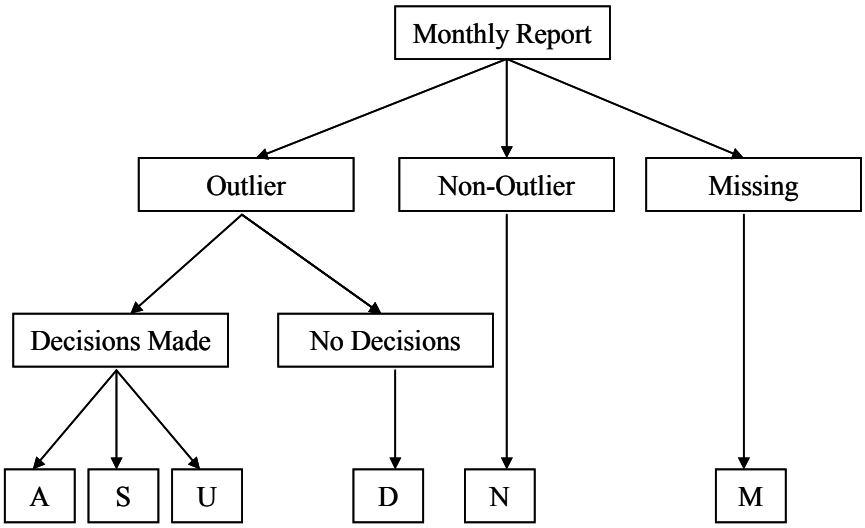
$E(\xi) = E(X) - E(Y) = \lambda - \mu = a - a = 0$ and
$V(\xi) = V(X) + V(Y) = \lambda + \mu = 2a$

are immediate in the above discussion. But, the explicit probability density function (A.1) of ξ is not immediate. The normality of the standardized form z in (A.2) arises from (5.2) under the condition $T = X_0 + Y_0$, by noting that

$$z_0 = \frac{2X_0 - T}{\sqrt{T}} = \frac{X_0 - Y_0}{\sqrt{X_0 + Y_0}}. \tag{A.4}$$

In testing the hypothesis $H_0 : \lambda = \mu$, it is reasonable to consider the standardized difference and reject the null hypothesis when the absolute value $|z|$ is too large.

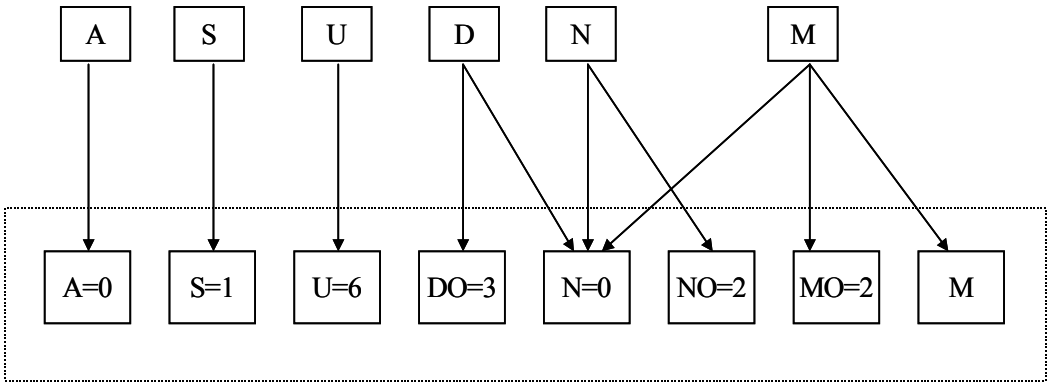
Appendix B
Ramifications of the Initial Monthly Offense Review



A = the outlier is accepted for full use by UCR.
S = the outlier is accepted for self-representation.
U = the outlier for the month is considered unacceptable.
D = the decision is deferred for the outlier.
N = the monthly report is not an outlier.
M = the report for the month is missing.

Appendix C
Ramifications of the Final Monthly Offense Review

The numerical values are assigned to represent the perceived severity of individual outcomes.

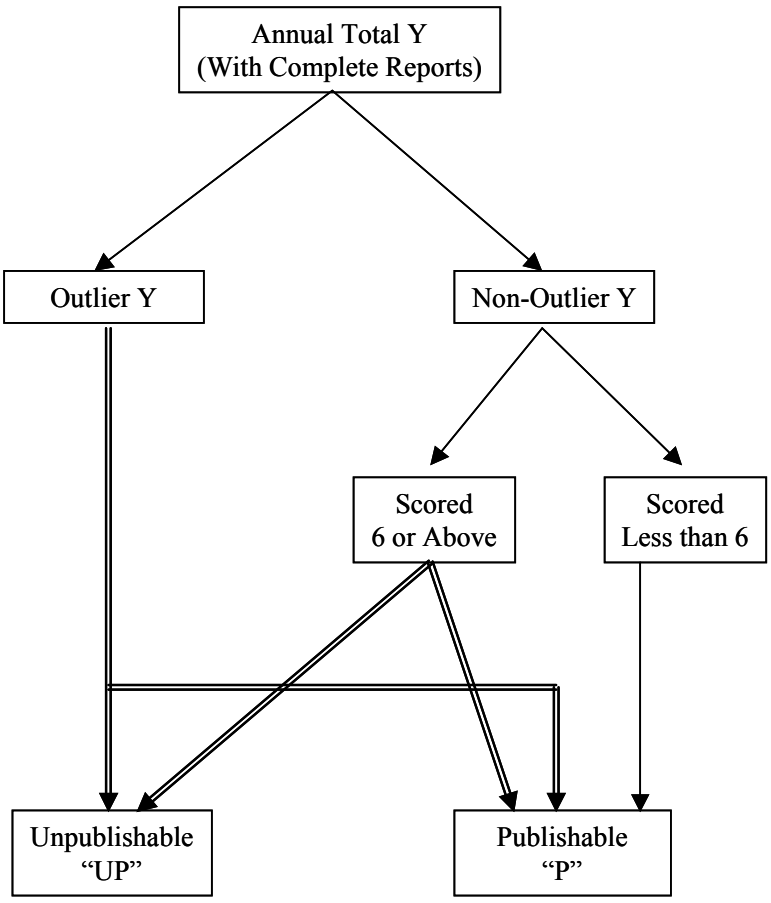


The above chart describes the outcome of the final monthly review for agencies that submit monthly reports. The codes in the first row (A, S, U, D, N, and M) represent the results of the initial monthly review, while those inside the larger box (the bottom row) are for the final monthly review. However, D in this appendix includes (a) outliers D for which no status decisions have been made (in the sense of Appendix B) and (b) human decisions A, S, D, and U that became outdated due to subsequent updating. Irrespective of the outlier status designated at the final monthly review, a report does not change the status indicator codes such as A, S, and U.

DO (= Deferred Outlier) denotes deferred-outlier (including outliers resulting from subsequent submissions). MO (= Missing Outlier) means a report was missing at the initial monthly review and later submitted for consideration resulting in an outlier at the final review. NO (= N became O) means that the initial report with N was later updated, and the updated report was an outlier.

Appendix D
Ramifications of the Annual Offense Review
(For Agencies that Submitted 12 Monthly Reports for a Given Data Category)

Double-lined arrows indicate human decisions, and single-lined arrows indicate decisions made by algorithms.



Appendix E
Rules of Monthly Data Estimations

Mode of Submission	Agency Status (1)	Data Category (2)	Annual Total (3)	Monthly Review (4)	Monthly Estimation (5)
Monthly	R	P			No
		UPP			Yes
	UR	P			No
		UP	O	U, DO, MO, NO	Yes
				N, A, S	No
			N	U	Yes
				N, A, S, DO, MO, NO	No
			I	M, U, DO, MO, NO	Yes
				N, A, S	No
		UPP			Yes
Other	R	P			No
		UPP			Yes
	UR	P			No
		UP, UPP			Yes

(1) R = Releasable agency.
UR = Unreleasable agency.

(2) P = Publishable crime category.
UP = Unpublishable crime category.
UPP= Unpublishable crime category due to policy.

(3) N = Annual total is a non-outlier.
O = Annual total is an outlier.
I = Incomplete crime category (due to the existence of missing months).

(4) A = The outlier is accepted for full use by UCR.
S = The outlier is accepted for self-representation.
U = The outlier for the month is considered unacceptable.
N = (a) The monthly report was not an outlier at the initial monthly review.
(b) The monthly report was missing at the initial monthly review, but the report submitted later was not an outlier at the final monthly review.
(c) The monthly report was an outlier at the initial monthly review, but human decision was deferred. It was no longer an outlier at the final monthly review.
M = The report for the month is still missing.

MO = Missing Outlier = the monthly report was missing at the initial monthly review but was submitted later. The report was an outlier at the final monthly review.

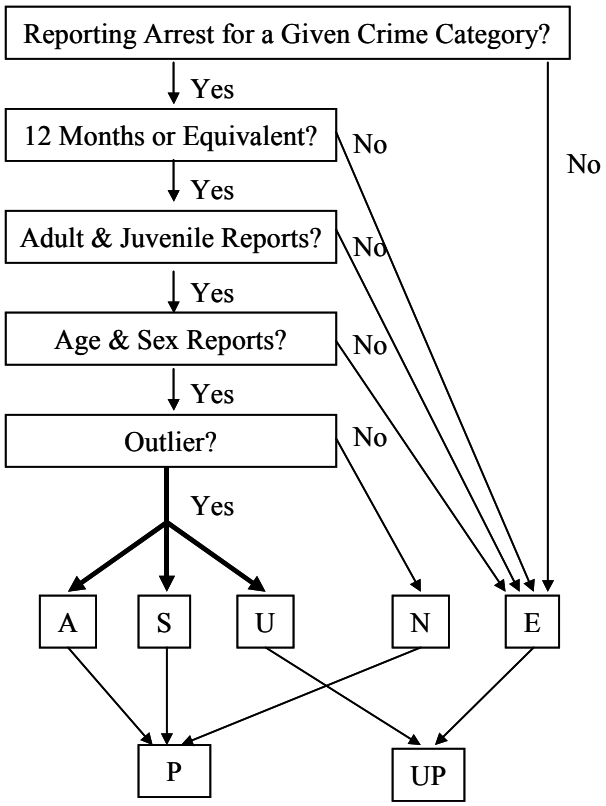
DO = Deferred-Outlier = the monthly report was an outlier at the initial monthly review, but the human decision was deferred or outdated due to new submission. At the final monthly review, it was still an outlier.

NO = The initial non-outlier report was later updated, and the new report was an outlier at the final monthly review.

(5) Yes = Estimation is made for the month.

No = No estimation is made for the month.

Appendix F
Annual Arrest Data Review



The heavy lines (arrows) represent human decisions. For a given crime category, the following notations are used.

N = The arrest total of the agency is not an outlier.
A = The arrest total of the agency is an outlier but was accepted for full use.
S = The arrest total of the agency is an outlier but was accepted for self-representation.
U = The arrest total of the agency is an outlier and unacceptable for use.
E = The agency was excluded from annual review. Its data will be estimated.
P = Publishable Crime Category.
UP = Unpublishable Crime Category.

References

1. Akiyama, Y. "Notes on UCR Statistics: Percent Change, Crime Rate, Crime Clock, Lifetime Chance for Murder Victimization, and Age-Specific Arrest Rate," Criminal Justice Information Services Division, Federal Bureau of Investigation, September 7, 2000.
2. Barnett, V. and Lewis, T. *Outliers in Statistical Data*, 3rd Edition, John Wiley & Sons, 1998.
3. Larsen, R. J. and Marx, M. L. *Introduction to Mathematical Statistics and Its Applications*, Prentice Hall, 1986.
4. Lehman, L. E. *Testing Statistical Hypothesis*, John Wiley & Sons, 1986.
5. Stuart, A. and Ord, J. K. *Kendall's Advanced Theory of Statistics*, Volume 1 (5^{th} Edition), Oxford University Press, 1987.
6. Stuart, A. and Ord, J. K. *Kendall's Advanced Theory of Statistics*, Volume 2 (5^{th} Edition), Oxford University Press, 1991.
7. Whitaker, Janet S. "ASUCRP Data Quality Panel," IBR Program, New York State: Prepared for the Annual Conference of the Association of State Uniform Crime Reporting Programs (ASUCRP), Vail, Colorado, October 11, 2000 (Revised June 2001).
8. Woolson, Robert F. *Statistical Method for the Analysis of Biomedical Data*, John Wiley & Sons, 1987.
9. Ryan, T. P. *Statistical Methods for Quality Improvement*, John Wiley & Sons, 1989.
10. Akiyama, Y. and Propheter, S. K. "Methods of UCR Data Estimation" (to appear), Criminal Justice Information Services Division, Federal Bureau of Investigation.