2-2023

Modeling and Fitting Two-Way Tables Containing Outliers

David L. Farnsworth
Rochester Institute of Technology

Recommended Citation
David L. Farnsworth, "Modeling and Fitting Two-Way Tables Containing Outliers", International Journal of Mathematics and Mathematical Sciences, vol. 2023, Article ID 6352058, 6 pages, 2023. https://doi.org/10.1155/2023/6352058

Hindawi
International Journal of Mathematics and Mathematical Sciences
Volume 2023, Article ID 6352058, 6 pages
https://doi.org/10.1155/2023/6352058

Research Article

Modeling and Fitting Two-Way Tables Containing Outliers

David L. Farnsworth

School of Mathematical Sciences, Rochester Institute of Technology, Rochester, NY 14623, USA

Correspondence should be addressed to David L. Farnsworth; dlfsma@rit.edu

Received 23 December 2022; Revised 29 January 2023; Accepted 30 January 2023; Published 11 February 2023

Academic Editor: Niansheng Tang

A model is proposed for two-way tables of measurement data containing outliers. The two independent variables are categorical and error-free. Neither missing values nor replication is present. The model consists of the sum of a customary additive part that can be fit using least squares and a part that is composed of outliers. Recommendations are made for methods for identifying cells containing outliers and fitting the model. A graph of the observations is used to determine the outliers' locations. For all cells containing an outlier, replacement values are determined simultaneously using a classical missing-data tool. The result is called the adjusted table. The inserted values are such that, when a mean-based fitting of the adjusted table is performed, the residuals in those cells are zero. The outlying portion of the observation in each of those cells is the difference of the observation and the replacement value. In this way, outliers are removed from further analyses of the adjusted table. This is particularly helpful because outliers can greatly contaminate and alter computations and conclusions. Subsequently, the causes of the outliers might be determined, and statistical estimation and testing can be implemented on the adjusted table.

1. Introduction

The modeling of an $m \times n$ table that contains outliers and is otherwise approximately additive is addressed. An outlier is an observation that is substantially different from the other observations ([1], p. 4; [2], p. 1; [3], pp. 3-4). Methods are recommended for the identification of the cells containing outliers, computation of replacement values for those cells, and estimation of the sizes of the outliers beyond the replacement values. They fill the need for a set of systematic and uncomplicated methods.

The goal is a least-squares, i.e., mean-based, analysis of the observations. However, when outliers are present, that statistical analysis is unreliable due to cross contamination among the table's cells. The cells that contain outliers are identified, and the outliers' sizes are estimated in order to determine whether they might be impactful on the statistical analysis. In the model and fitting procedure, the underlying additive observations are prepared for analyses, and the outlying measurements, being isolated both in the model and the fit, can be concurrently investigated, including for causes.

The approach is guided by a desire for simplicity, believing that little is gained from excessive computations and manipulation of the observations and that, indeed, much can be lost with subsequent misleading outcomes. Imposing or guessing a probability model can cause its own problems. If it is incorrect, non-outlying observations might be misidentified as outlying because they do not fit the model or might be missed because they do ([2], pp. 60–62).

It is required that the outliers' identification method always works, that is, outliers are not missed and non-outliers are not identified as outliers. The method should not presuppose a certain fixed number of outliers, as some methods do. A governing principle is that the methods must not break down.

The cells that contain outlying values are found by examining a three-dimensional graph of the observations, where it is extremely unlikely that outliers will be masked or non-outliers will be misidentified as outliers. Each outlier is removed from the analysis, and a perfectly fitting value is inserted into its cell. The cell's replacement value eliminates repercussions by the outlier on further analyses. The outliers themselves can be important in many ways ([2], pp. 399–418;

[3], pp. 16-17, 199-200). To the person using the observations, the role of the outlying observations can be either beneficial or harmful. The outliers might indicate a cure for a disease or combinations of the independent variables that produce uniquely favorable outcomes with either high or low values of the dependent variable. On the other hand, outliers may highlight pairs of values of the independent variables that cause harm and should be avoided. The identified outlying values can be errors and might be investigated as such.

For measurement data y_{ij}, the model is

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} + \sum_{(p,q)} \delta_{ij} I_{(i,j)=(p,q)}, \tag{1}$$

with $1 \le i \le m$, $1 \le j \le n$, $m \ge 3$, and $n \ge 3$, where there is one observation per cell. The first four terms are considered a conventional additive, mean-based model. Cells identified as (p, q) contain outliers of size δ_{pq} beyond the additive model. The function I is the indicator function.

The graph of the observations serves the purpose of finding the locations of the outliers and checking the appropriateness of the additive model. The independent axes are the row and column numbers i and j. The graphing program allows rotation of the axes, so that the observations can be viewed from any vantage point. If the observations do not appear sufficiently planar, a transformation of the dependent variable should be considered. If the points corresponding to a whole row or column are questionable, that level of the dependent variable should be examined.

To simultaneously determine replacement values for all the cells identified as having an outlier, place a variable in each of those cells and find the additive fit for them. The fitted values depend upon the variables in a linear way. Set each fitted value equal to the respective variable. The solution of this system of linear equations is the replacement value for the adjusted table. They give residuals of zero in those cells in mean-based fitting of the adjusted table. The estimates of the sizes of the outlying portions are the differences between the observations and the replacement values. The outlying portions of the cells' observations are segregated from the adjusted table. For each replaced value, one degree of freedom is lost in subsequent analyses of the $m \times n$ adjusted table.

Example 1. Consider the artificial data in Table 1 [4]. Figure 1 shows that, except for the two outliers in cells (1, 1) and (3, 3), the data appear to be close to planar, thus additive. No transformation is required. Placing y'_{11} and y'_{33} in cells (1, 1) and (3, 3), finding the fitted values, and setting the fitted values equal to y'_{11} and y'_{33} gives

$$\frac{y'_{11} + 2 + 1 + 2}{4} + \frac{y'_{11} + 2 + 2}{3} - \frac{y'_{11} + 2 + 1 + 2 + 2 + 0 + 2 + 2 + 2 + 1 + y'_{33} + 0}{12} = y'_{11}, \tag{2}$$

and

$$\frac{2 + 1 + y'_{33} + 0}{4} + \frac{1 + 2 + y'_{33}}{3} - \frac{y'_{11} + 2 + 1 + 2 + 2 + 0 + 2 + 2 + 2 + 1 + y'_{33} + 0}{12} = y'_{33}, \tag{3}$$

or $6y'_{11} + y'_{33} = 17$ and $y'_{11} + 6y'_{33} = 7$, whose solution gives the replacement values $y'_{11} = 19/7$ and $y'_{33} = 5/7$. Fitting the adjusted table containing these values and using the outlying portions $14–19/7 = 79/7$ and $5–5/7 = 30/7$ in cells (1, 1) and (3, 3), the fit of the model is

$$y_{ij} = \frac{61}{42} + \begin{pmatrix} \frac{10}{21} \\ \\ \frac{1}{21} \\ \\ -\frac{11}{21} \end{pmatrix} + \left(\frac{33}{42} - \frac{19}{42} - \frac{9}{42} - \frac{5}{42}\right) + \frac{79}{7} I_{(i,j)=(1,1)} + \frac{30}{7} I_{(i,j)=(3,3)}. \tag{4}$$

2. Computing the Replacement Values

This section formally presents the method employed to compute the values for insertion into the adjusted table to replace the values in the cells containing outliers. The method's advantages are that it is computationally simple even for a large number of outliers and requires no iterations. The values are unique, if no entire row or column is considered to be outlying. The values yield residuals of zero in subsequent mean-based analyses of the adjusted table.

Designate the original $m \times n$ table by T. Table T' is the same as T, except that the cells containing outliers have had their observations replaced by placeholders. The placeholder in cell (h, k) is designated by y'_{hk}. Table T'' is the same as T, except that the cells of T containing outliers are composed of zeros. The fitted value for cell (h, k) in T is

$$f_{hk} = \frac{\sum_{j=1}^{n} y_{hj}}{n} + \frac{\sum_{i=1}^{m} y_{ik}}{m} - \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} y_{ij}}{mn}. \tag{5}$$

The fitted values for cell (h, k) in T' and T'' are designated as f'_{hk} and f''_{hk}, respectively.

Table 1: Tukey's artificial data for Example 1.

14	2	1	2
2	0	2	2
2	1	5	0

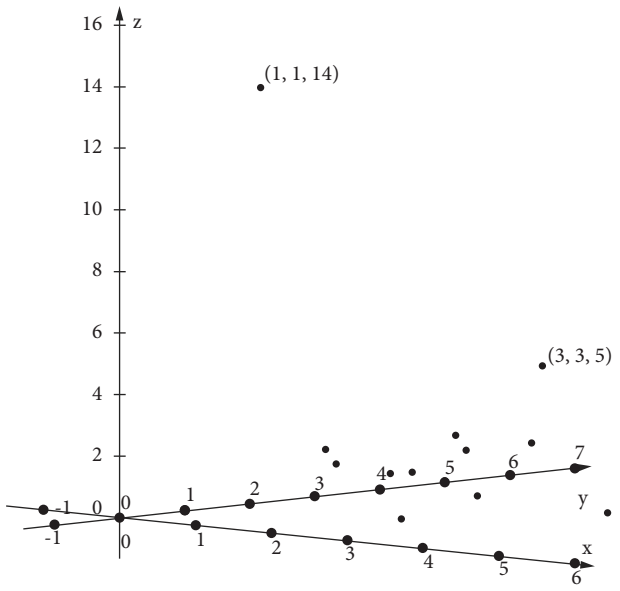

Figure 1: Observations in Table 1 with $(x, y, z) = (i, j, y_{ij})$.

Theorem 1 (one outlier). *If T has one outlier, which is in cell (h, k), then the replacement value is*

$$y'_{hk} = \frac{mn}{(m-1)(n-1)} f''_{hk}. \tag{6}$$

Proof. Using (2) and setting the fitted value in T' equal to y'_{hk} gives

$$f'_{hk} = \frac{y'_{hk}}{n} + \frac{y'_{hk}}{m} - \frac{y'_{hk}}{mn} + f''_{hk} = y'_{hk}, \tag{7}$$

whose solution is (6).

The outlying portion of the observation in cell (h, k) in Theorem 1 is defined to be $y_{hk} - y'_{hk}$, which is an estimator of δ_{hk} in model (1). □

Theorem 2 (two outliers). *If T has two outliers, which are in cells (h, k) and (s, t), then the replacement values are given by the solutions to the following systems of linear equations, whose solutions are unique.*

Case 1 (outliers in different rows and columns). If $s \neq h$ and $t \neq k$, then

$$(m-1)(n-1)\, y'_{hk} + y'_{st} = mn f''_{hk}, \tag{8}$$

and

$$y'_{hk} + (m-1)(n-1)\, y'_{st} = mn f''_{st}. \tag{9}$$

Case 2 (outliers in the same row and different columns). If $s = h$ and $t \neq k$, then

$$(m-1)(n-1)\, y'_{hk} - (m-1)\, y'_{ht} = mn f''_{hk}, \tag{10}$$

and

$$-(m-1)\, y'_{hk} + (m-1)(n-1)\, y'_{ht} = mn f''_{ht}. \tag{11}$$

Case 3 (outliers in different rows and the same column). If $s \neq h$ and $t = k$, then

$$(m-1)(n-1)\, y'_{hk} - (n-1)\, y'_{sk} = mn f''_{hk}, \tag{12}$$

and

$$-(n-1)\, y'_{hk} + (m-1)(n-1)\, y'_{sk} = mn f''_{sk}. \tag{13}$$

Proof. For Case 1 using (5),

$$f'_{hk} = \frac{y'_{hk}}{n} + \frac{y'_{hk}}{m} - \frac{y'_{hk} + y'_{st}}{mn} + f''_{hk} = y'_{hk}, \tag{14}$$

and

$$f'_{st} = \frac{y'_{st}}{n} + \frac{y'_{st}}{m} - \frac{y'_{hk} + y'_{st}}{mn} + f''_{st} = y'_{st}, \tag{15}$$

which are equivalent to (9). Since the determinant of the coefficients is $(m-1)^2(n-1)^2 - 1 > 0$, there is just one solution.

For Case 2,

$$f'_{hk} = \frac{y'_{hk} + y'_{ht}}{n} + \frac{y'_{hk}}{m} - \frac{y'_{hk} + y'_{ht}}{mn} + f''_{hk} = y'_{hk}, \tag{16}$$

and

$$f'_{ht} = \frac{y'_{hk} + y'_{ht}}{n} + \frac{y'_{hk}}{m} - \frac{y'_{hk} + y'_{ht}}{mn} + f''_{ht} = y'_{ht}, \tag{17}$$

which yields (11), where the determinant of the coefficients is non-zero.

The proof of Case 3 is similar to the proof of Case 2.

The following theorem can be proven with the method of the proof of Theorem 2. □

Theorem 3 (three outliers). *If T has three outliers, which are in cells (h, k), (s, t), and (u, v), then the replacement values are given by the solutions to the following systems of linear equations.*

Case 4 (outliers in different rows and columns). If k, s, and u are distinct and k, t, and v are distinct, then

$$\begin{aligned} (m-1)(n-1)\, y'_{hk} + y'_{st} + y'_{uv} &= mn f''_{hk}, \\ y'_{hk} + (m-1)(n-1) y'_{st} + y'_{uv} &= mn f''_{st}, \end{aligned} \tag{18}$$

and

$$y'_{hk} + y'_{st} + (m-1)(n-1)\, y'_{uv} = mn f''_{uv}. \tag{19}$$

Case 5 (two outliers in one row and the third in a different row and all outliers are in different columns). If $s = h \neq u$ and k, t, and v are distinct, then

$$\begin{aligned}(m-1)(n-1)y'_{hk} - (m-1)y'_{ht} + y'_{uv} &= mnf''_{hk},\\ -(m-1)y'_{hk} + (m-1)(n-1)y'_{ht} + y'_{uv} &= mnf''_{ht},\end{aligned} \tag{20}$$

and

$$y'_{hk} + y'_{ht} + (m-1)(n-1)y'_{uv} = mnf''_{uv}. \tag{21}$$

Case 6 (two outliers in one column and the third in a different column and all outliers are in different rows). If $t = k \neq v$ and h, s, and u are distinct, then

$$\begin{aligned}(m-1)(n-1)y'_{hk} - (n-1)y'_{sk} + y'_{uv} &= mnf''_{hk},\\ -(n-1)y'_{hk} + (m-1)(n-1)y'_{sk} + y'_{uv} &= mnf''_{sk},\end{aligned} \tag{22}$$

and

$$y'_{hk} + y'_{sk} + (m-1)(n-1)y'_{uv} = mnf''_{uv}. \tag{23}$$

Case 7 (two outliers in one row and the third in the same column as one of the two). If $s = h \neq u$ and $v = k \neq t$, then

$$\begin{aligned}(m-1)(n-1)y'_{hk} - (m-1)y'_{ht} - (n-1)y'_{uk} &= mnf''_{hk},\\ -(m-1)y'_{hk} + (m-1)(n-1)y'_{ht} + y'_{uk} &= mnf''_{ht},\end{aligned} \tag{24}$$

and

$$-(n-1)y'_{hk} + y'_{ht} + (m-1)(n-1)y'_{uk} = mnf''_{uk}. \tag{25}$$

Case 8 (outliers in one row). If $h = s = u$ and k, t, and v are distinct, then

$$\begin{aligned}(m-1)(n-1)y'_{hk} - (m-1)y'_{ht} - (m-1)y'_{hv} &= mnf''_{hk},\\ -(m-1)y'_{hk} + (m-1)(n-1)y'_{ht} - (m-1)y'_{hv} &= mnf''_{ht},\end{aligned} \tag{26}$$

and

$$-(m-1)y'_{hk} - (m-1)y'_{ht} + (m-1)(n-1)y'_{hv} = mnf''_{hv}. \tag{27}$$

Case 9 (outliers in one column). If $k = t = v$ and h, s, and u are distinct, then

$$\begin{aligned}(m-1)(n-1)y'_{hk} - (n-1)y'_{sk} - (n-1)y'_{uk} &= mnf''_{hk},\\ -(n-1)y'_{hk} + (m-1)(n-1)y'_{sk} - (n-1)y'_{uk} &= mnf''_{sk},\end{aligned} \tag{28}$$

and

$$-(n-1)y'_{hk} - (n-1)y'_{sk} + (m-1)(n-1)y'_{uk} = mnf''_{uk}. \tag{29}$$

In Cases 4–7 of Theorem 3, the determinants of the coefficients of the equations are non-zero for $m \geq 3$ and $n \geq 3$, so there is a unique solution. In Case 8, the determinant is $-(m-1)^3 n(n-3)$, which is zero solely for $n = 3$ columns. If the table has only three columns and a particular row has three outliers, the estimates are not uniquely determined. Case 9 is similar with more than three rows being required for uniqueness of the estimates. Cases 8 and 9 illustrate the requirement for a unique solution that no single row or column is comprised of outliers.

A pattern emerges from Theorems 1–3.

Theorem 4 (any number of outliers). *For v cells that contain outliers, the replacement values are given by the solutions to the system of linear equations:*

$$M_{UV} Y'_V = mnF''_U, \tag{30}$$

which are given in matrix notation, where the subscripts are for all the cells (i, j) that are among the cells (p, q), as displayed in model (1). The vector Y'_V is a column of the replacement values y'_{ij}, which can be listed in any order, as long as consistency is maintained. The elements of column vector F''_U are the appropriate f''_{ij}. The $v \times v$ matrix M_{UV} is symmetric. Each element on the main diagonal is $(m-1)(n-1)$. The off-diagonal elements are $+1$, $-(m-1)$, or $-(n-1)$. The entry depends upon whether U and V represent cells that are in neither the same row nor the same column, in the same row, or in the same column, respectively. The solution is unique if there is no complete row or column of outliers.

If all cells of a row or column are deemed to contain outliers, then the solution is not unique because adding any number to each replacement value is also a solution. At least one non-outlying cell is needed in the row or column in order to anchor the row or column's run, so that the remaining values are unique. As few as minimum $\{m, n\}$ outliers are sufficient for non-uniqueness. A row or column of outliers would indicate that the corresponding level of the independent variable should be examined before further analyses are attempted, so this is not a threat to this procedure.

The other extreme is that all replacement values are unique if the cells are in an $m-1$ by $n-1$ block. In that case, the determinant of $M_{UV} \neq 0$, and a perfectly fitting, additive adjusted table is obtained. The row outside the block determines the differences for the additivity of the replacement values, and the single data value in each of the other rows determines the scale of its row. Because one degree of freedom is lost in the error term of (1) for each replacement value, determining the block's $(m-1)$ $(n-1)$ replacement values reduces the degrees of freedom to zero. This maximum number of outliers would indicate that the so-called outliers are the typical data. The number of cells that might be considered to contain outliers gives reassurance that the breakdown of this method will not be met in practice. Indeed, this method is very safe.

3. Example

Example 2. Consider the data in Table 2. A graph shows that there are high outliers in cells (1, 2) and (1, 3) and a low outlier in cell (3, 4). This is an instance of Case 5 of Theorem 3 with $(h, k) = (1, 2)$, $(h, t) = (1, 3)$, and $(u, v) = (3, 4)$. Table T'' is given in Table 3.

The equations for the replacement values are

Table 2: Artificial data T for Example 2.

1	10	12	4	5
6	7	8	9	10
11	12	13	4	15

Table 3: T'' corresponding to T in Table 2.

1	0	0	4	5
6	7	8	9	10
11	12	13	0	15

Table 4: Artificial observations to show the impact of an outlier.

1	2	3
4	5	6
7	8	9

Table 5: Residuals for an additive fit to Table 4 with 9 added to the cell (1, 3).

−2	−2	4
1	1	−2
1	1	−2

$$\begin{aligned} 8y'_{12} - 2y'_{13} + y'_{34} &= 24, \\ -2y'_{12} + 8y'_{13} + y'_{34} &= 34, \end{aligned} \tag{31}$$

and

$$y'_{12} + y'_{13} + 8y'_{34} = 117, \tag{32}$$

whose solution is $y'_{12} = 2$, $y'_{13} = 3$, and $y'_{34} = 14$ The corresponding outlying portions in those cells are $10 - 2 = 8$, $12 - 3 = 9$, and $4 - 14 = -10$.

4. Concluding Comments

The cells designated as (p, q) in (1) are few in number and contain values that might be called outliers, interactions, errors, blunders, or contaminations or described as renegade, rogue, spurious, deviant, unrepresentative, stray, discordant, incongruous, or wild. The designation outlier is used in order to avoid words that might appear to prejudice the search for the cause of an unusual value. The model clarifies the manner in which outliers are set aside for examination.

The graphical method for identifying outliers does not involve any manipulation of the observations, and thus no assumptions or artifacts of calculations enter the procedure. In particular, it does not involve residuals, which are widely used but fraught with pitfalls ([1], pp. 281–328). An aim is to alter the observations as little as possible by using a very light touch.

A graph of the observations may be the safest method for determining the locations of outliers. Mean-based methods using z-scores can be flawed because outliers tend to increase standard deviations so that z-scores are reduced [5, 6]. Mean-based procedures have the feature that outliers contaminate multiple cells and lose their impact on their own cells ([1], pp. 284–285, 301–318; [7], pp. 184–185; [8]). An outlier biases the statistics that are being used to detect it and introduces dependence among the residuals. For a very simple example, take the artificial observations in Table 4, which have residuals of zero for an additive fit. Adding 9 to the entry in cell (1, 3) and performing a mean-based additive fit produce the residuals in Table 5. The spurious value in cell (1, 3) has produced non-zero residuals in every cell. Although the largest residual is in cell (1, 3), if some noise was present, the outlier could avoid detection and raise the variance for a subsequent estimation or testing procedure.

Graphing allows the observations to speak for themselves with no contamination or distortions and, importantly, with no assumption about the number of outliers. Decisions about the existence of a possible outlier do not have to be made through the lens of statistical calculations. Many graphing programs are freeware and work on many different platforms. They allow the user to rotate the axes and thus to examine the data cloud from any viewpoint, so that observations away from the data cloud can be seen and identified. One program is GeoGebra [9] (available at https://www.geogebra.org), which is used for Figure 1.

Some other options for identifying outliers are median based because the median is known for its resistance to outliers. Two options that are often suggested are to fit the table using median polish ([7], pp. 184–185) or to determine whether observation is outside the inner fences, as in a boxplot [10]. However, both options have weaknesses. Median polish has the significant disadvantage that it can fail when there are outliers in all but one of the entries of a row or a column [8]. So, for a 3×3 table with outliers in cells (1, 1) and (1, 2), median polish might indicate that there is an outlier in cell (1, 3). Other unsatisfactory features of median polish are that different results might be obtained depending upon whether rows or columns are polished first and whether the mid-median or the low median is used. The procedure can fail to terminate and may even cycle.

The inner fences are $Q_1 - 1.5$ IQR and $Q_3 + 1.5$ IQR, where Q_1 and Q_3 are the first and third quartiles and IQR is the interquartile range $Q_3 - Q_1$. Although this tool can be effective for univariate data ([2], pp. 45-46; [3], pp. 13-14), it ignores the row-column structure of the observations. It fails for tables that are nearly additive and tilted. For example, if the entries increase greatly to the right, and the outlier occurs at the cell in the first column whose underlying values are small, then the outlier might be about the size of the values in the right-hand columns and be undetected in the undifferentiated batch of observations. It would be discovered in a graph, as it would be far from the data cloud.

There exist many diverse strategies for identifying outlying observations, which involve extra assumptions, sums of squares or other distance measures or even altering the definition of outlying to being on the boundary of a data cloud, all of which we have sought to circumvent. They include looking for reduction in variance [11], nearest neighbor techniques [12], and finding edges and corners in the data cloud [13]. Another technique is to search for low-density areas in the cloud of data [14], which can use few assumptions but can be computationally intensive. Some of these methods are difficult to implement when the independent variables' values are not ordered, which is the case addressed here.

The process used for finding the replacement values in the adjusted table reflects the operations based on means, which are envisioned for the adjusted table; consequently, no new concepts are added to the analysis. The calculations are extremely simple; in particular, no iterations are involved.

The replacement values are numerically the same as Yates' values used in cells where there are missing values ([15]; [16], pp. 33–34). For just one missing value, which is in cell (h, k),

$$y'_{hk} = \frac{m\sum_{j=1}^{n} y_{hj}'' + n\sum_{i=1}^{m} y_{ik}'' - \sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij}''}{(m-1)(n-1)}, \tag{33}$$

has been suggested to be inserted into the cell with the missing value ([16], p. 34; [17]; [18], p. 9). Expressions (6) and (33) are equivalent. Adding and subtracting y_{hk} in each of the three summations in (33) yields

$$y'_{hk} = \frac{mnf_{hk} + (1-m-n)y_{hk}}{(m-1)(n-1)}, \tag{34}$$

which is a weighted average that does not depend upon the value of y_{hk}. (34) uses quantities that contain y_{hk} and thereby makes unnecessary the introduction of T' and T''.

There may be arguments in favor of other replacement values. Leaving a hole is not recommended ([8], [16], pp. 4, 32–33). A candidate for a replacement value for a single outlier in cell (h, k) is a weighted average of values in the cell's row and column. Using a cell's nearest neighbors requires care because neither the rows nor the columns are ordered. The only neighbors to a cell are the cells in its row and column, and all those values are equally close.

When there are multiple outliers, replacement values for the cells are determined simultaneously, so that the cells' residuals would be zero in a mean-based fit of the table with the replacement values. In this way, each cell's value does not contaminate the replacement values for the other cells. A sequential accommodation of outliers that might eliminate them one at a time has that shortcoming.

The scope of this paper is limited to categorical or qualitative independent variables and quantitative dependent variables. Independent variables with more structure, such as an ordering, might be taken advantage of by using the fact that it creates a proximity measure. If the dependent variables are counts with a sufficient range, they might be treated as if they were continuous, instead of discrete. The field of outlier detection and analysis for categorical dependent variables is large and very active ([2], pp. 249–272; [3], pp. 69–93). Its techniques depend upon the features of the particular variables, in order to define nearness or similarities and to identify the appropriate discrete probability distributions.

The suggested methodology is practical. Users are armed with knowing the locations of the irregular observations and possessing estimates of the sizes of these unusual values. They can decide if they are meaningfully large in magnitude and whether or not to search for causes. Since the recommended technique is very accessible, straightforward, and easy to implement, this is a suitable topic to introduce into elementary statistics courses.

Data Availability

No data were used to support this study.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] V. Barnett and T. Lewis, *Outliers in Statistical Data*, Wiley, New York, NY, USA, 2nd edition, 1984.

[2] C. C. Aggarwal, *Outlier Analysis*, Springer Nature, Berlin, Germany, 2nd edition, 2017.

[3] N. N. R. R. Sui, M. N. Murty, and G. Athithan, *Outlier Detection: Techniques and Applications*, Springer, Berlin, Germany, 2019.

[4] J. W. Tukey, "One degree of freedom for non-additivity," *Biometrics*, vol. 5, no. 3, pp. 232–242, 1949.

[5] D. C. Hoaglin, "Using leverage and influence to introduce regression diagnostics," *The College Mathematics Journal*, vol. 19, no. 5, pp. 387–416, 1988.

[6] R. E. Shiffler, "Maximum Z scores and outliers," *The American Statistician*, vol. 42, no. 1, pp. 79-80, 1988.

[7] F. Mosteller and J. W. Tukey, *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley, Boston, MA, USA, 1977.

[8] J. D. Emerson and D. C. Hoaglin, "Analysis of two-way tables by medians," in *Understanding Robust and Exploratory Data Analysis*, D. C. Hoaglin, F. Mosteller, and J. W. Tukey, Eds., Wiley, New York, NY, USA, 1983.

[9] Geogebra, "GeoGebra for teaching and learning math," 2022, https://www.geogebra.org.

[10] C. Goodall, "Examining residuals," in *Understanding Robust and Exploratory Data Analysis*, D. C. Hoaglin, F. Mosteller, and J. W. Tukey, Eds., Wiley, New York, NY, USA, 1983.

[11] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos, "LOCI: fast outlier detection using the local correlation integral," in *Proceedings of the 19th International Conference on Data Engineering, IEEE Computer Society*, pp. 315–326, Bangalore, India, March 2003.

[12] J. Zhang and H. Wang, "Detecting outlying subspaces for high-dimensional data: the new task, algorithms and performance," *Knowledge and Information Systems*, vol. 10, no. 3, pp. 333–355, 2006.

[13] I. Ruts and P. J. Rousseeuw, "Computing depth contours of bivariate point clouds," *Computational Statistics and Data Analysis*, vol. 23, no. 1, pp. 153–168, 1996.

[14] C. C. Aggarwal and P. S. Yu, "Outlier detection in high dimensional data," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 37–46, Santa Barbara, CA, USA, May 2001.

[15] F. Yates, "The analysis of the replicated experiments when the field results are incomplete," *Empire Journal of Experimental Agriculture*, vol. 1, pp. 129–142, 1933.

[16] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, Wiley, New York, NY, USA, 3rd edition, 2020.

[17] F. E. Allan and J. Wishart, "A method of estimating the yield of a missing plot in field experimental work," *The Journal of Agricultural Science*, vol. 20, no. 3, pp. 399–406, 1930.

[18] W. G. Cochran and G. M. Cox, *Experimental Designs*, Wiley, 1950.