9-2024

Analysis of r x c Tables containing Outliers

David L. Farnsworth
Rochester Institute of Technology

Journal of Probability and Statistical Science
22(1), 1-15 Sep. 2024

Analysis of $r \times c$ Tables containing Outliers

David L. Farnsworth
School of Mathematics and Statistics
Rochester Institute of Technology, Rochester, New York, USA

ABSTRACT

We examine the problem of accommodating outliers in $r \times c$ tables of measurement data when there are no replications. The suggested strategy for the determination of the cells that contain outliers and for handling them analytically is easily understood and implemented in practice and presented in even the most elementary statistics course along with the course's ANOVA material.

1. Introduction

Outliers are persistent problems in statistics. They are characterized as being far from the other observations [1, 2, 8]. For the univariate set {2,2,2,12}, where it is believed that the 12 is in error, the sample mean 4 is double the typical value 2. The 12 has a huge influence on the mean. One remedy is to use the median, which is 2. Often, these kinds of examples are shown in class. As the observations' structure becomes more complicated, identifying and dampening the influence of outlying values becomes more difficult. We present a way to do that for two-factor tables of measurement data with no replications. The goal is to show how the method, which is in Sections 7 and 8, can be presented in classes. The method is easily understood and implemented, even in the most elementary statistics course, perhaps being presented along with the course's ANOVA material.

The straightforward examples of Tables 1 and 3 are the main ones that we employ to illustrate the ideas. Table 3 was created by adding 18 to the entry in cell $(1,1)$ of Table 1, thereby causing an outlier to appear there. That analysis is in Sections 2–7. Another example is presented in Section 8. For pedagogical reasons, the tables contain integers and the factors' levels are arranged for visual simplicity. Especially for computer-lab sessions, we favor sets of real data, but here more instructional clarity is reached when the observations are integers.

The notation for mean-based additive fitting of a table is:

The observations, that is, the table's entries, are x_{ij} for $i = 1, 2, \dots, r$ and $j = 1, 2, \dots, c$.

The *overall effect* E is the mean or numerical average of all the table's entries.

The *row effect* R_i for $i = 1, 2, \dots, r$ is the i^{th} row mean minus the overall effect, that is, $R_i = \sum_{j=1}^{c} x_{ij}/c - E$.

The *column effect* C_j for $j = 1, 2, \dots, c$ is the j^{th} column mean minus the overall effect, that is, $C_j = \sum_{i=1}^{r} x_{ij}/r - E$.

The *fitted value* for cell (i, j) is $f_{ij} = E + R_i + C_j$.

The *residual* for cell (i, j) is $r_{ij} = x_{ij} - f_{ij}$.

Received March 2024, in final form August 2024.
David L. Farnsworth (corresponding author), is affiliated with the School of Mathematics and Statistics, Rochester Institute of Technology, Rochester, New York, USA
dlfsma@rit.edu

Each table entry is *additively decomposed* as

$$x_{ij} = f_{ij} + r_{ij} = E + R_i + C_j + r_{ij}.$$

Additive fitting is performed for Table 1 in Table 2. Table 1 is seen to be exactly additive, because the fitted values are the same as the original entries, that is, the residuals are all zero.

The analysis is in two parts. First, outliers need to be identified. Second, if there are outliers, they should be removed or neutralized, so that they do not contaminate the table's fitting or statistical analyses, such as ANOVA. A deleterious effect of an outlier's contamination is displayed in Section 4.

Table 3 is employed as the main example to introduce the graphical display in Section 2 and to discuss and compare analyses methods for tables with outliers, which appear in Sections 3–7. Another table, which contains two outliers, is presented in Section 8. The relatively new and recommend method for the analysis the outliers, which is our Method 4, is introduced in Section 7. The conclusions are in Section 9.

Table 1. An exactly additive table

x_{ij}		Second Factor		
		1	**2**	**3**
First Factor	**1**	16	11	9
	2	12	7	5
	3	8	3	1

Table 2. Additive fitting of Table 1 is performed

		Second Factor			Row Mean	Row Effect, R_i
		1	**2**	**3**		
First Factor	**1**	16	11	9	12	4
	2	12	7	5	8	0
	3	8	3	1	4	–4
Column	**Mean**	12	7	5	**Overall Mean** = 8	
Column	**Effect, C_j**	4	–1	–3		**Overall Effect, E** = 8

Table 3. An outlier has been introduced into cell (1,1) of Table 1 by adding 18

x_{ij}		Second Factor		
		1	**2**	**3**
First Factor	**1**	34	11	9
	2	12	7	5
	3	8	3	1

2. Identifying outliers graphically

In order to identify outliers, if any, we recommend using a graph of the rc points (i, j, x_{ij}). The graph should be rotatable in order to see the points, so that points cannot be masked, that is, hidden, by other points. At the same time, the structure of all the observations can be examined to determine whether some procedure is required, such a transformation of the dependent variable that yields the entries, to make the observations closer to planar for the suitability of linear fitting and statistical analyses. This graphical technique has the advantage that the user is looking directly at the observations, instead of a set of derived numbers that might hide outliers or falsely claim that some non-outlying observations are outliers [1]. Additionally, assumptions that may not be necessary or warranted can be embedded in formulas.

Figure 1 contains the diagnostic graph for Table 3. Clearly, there is just one outlier, which is at $(1, 1, 34)$, representing cell $(1,1)$. The other points appear to be close to a plane.

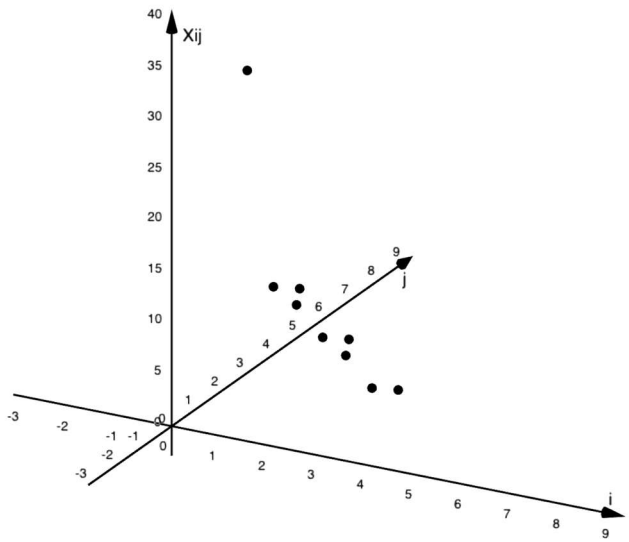

Figure 1. Diagnostic graph for Table 3, showing one outlier and planarity otherwise

3. Analysis of outlying observations

We want a procedure for finding an $r \times c$ table's fitted values f_{ij} with $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$, such that the fitted values will accommodate follow-up least-squares analyses and, in the event of some outliers, isolate the discordant portion of each outlier, perhaps for separate study. Also, we want the procedure to be able to handle very many suspect observations and to alert the user in the event that the procedure is breaking down, so that spurious results are not produced.

Consider four methods for fitting observations that are in a two-factor table with outliers, which are

1. leave known outliers in the table,
2. omit cells that contain outliers,
3. use medians via median polish, and
4. use the new suggested method.

In Sections 4–7, we analyze Table 3 with these four methods to compare them. Method 1 does a poor job, because the outlying portion in cell $(1,1)$ corrupts the procedure in all the cells, which illustrates the importance of identifying and removing outliers before fitting or performing other analyses, such as ANOVA [4, 7]. Performing Method 2 illustrates that it leads to fitted values and residuals that have no clear meanings, so that method is not advised [4, 6]. The third method, median polish, can be an excellent process for very few outliers, but fails without warning if half or more of the entries in any row or column are outliers, as those values are used in the procedure to compute medians [4, 5, 9]. The fourth method is relatively new and recommended [3]. In Section 8, another numerical example is analyzed with Method 4.

4. Method 1: Leave outliers in the table

Table 4 displays the fitting procedure for Table 3. The fitted values are in Table 5, and the residuals are in Table 6. The main problem is that the outlying value 34 in cell $(1,1)$ contaminates all of the fitted values, which can be seen from the residuals in Table 6 being nonzero. There is a pattern in the residuals in Table 6, but, unfortunately, we cannot take advantage of it, because any pattern would be camouflaged when there are random fluctuations in the observed values, which real data would have.

In Tables 2 and 4, the row effects and the column effects add to zero, and in Table 6, each row and each column of residuals add to zero. These zeros always occur when fitting with means, when there are no empty cells.

Table 4. Mean based fitting of Table 3, which contains an outlier in cell $(1,1)$

		Second Factor			**Row**	**Row**
		1	**2**	**3**	**Mean**	**Effect, R_i**
First Factor	**1**	34	11	9	18	8
	2	12	7	5	8	–2
	3	8	3	1	4	–6
Column	**Mean**	18	7	5	**Overall Mean** = 10	
Column	**Effect, C_j**	8	–3	–5		**Overall Effect, E** = 10

Table 5. Mean-based fitted values $f_{ij} = E + R_i + C_j$ for Table 4

f_{ij}		Second Factor		
		1	**2**	**3**
First Factor	**1**	26	15	13
	2	16	5	3
	3	12	1	–1

Table 6. Mean-based residual values $r_{ij} = x_{ij} - f_{ij}$ from Tables 3 and 5

r_{ij}		Second Factor		
		1	**2**	**3**
First Factor	**1**	8	–4	–4
	2	–4	2	2
	3	–4	2	2

5. Method 2: Omit cells containing outliers

In this section, we omit the value 34 in cell (1,1) of Table 3 and move forward with the fitting with means, which is performed in Table 7. Omitting the value unbalances the table and leaves no clear information in the fitted values and residuals, which are in Tables 8 and 9. The imbalance is demonstrated by the row effects and the column effects not adding to zero in Table 7 and by the rows and columns of the residuals not adding to zero in Table 9, unlike the corresponding sums in Section 4. A potential estimate for f_{11} is $E + R_1 + C_1 = 7 + 3 + 3 = 13$ from Table 7, but its meaning and veracity are doubtful [4, 6].

Table 7. Table 3 is fitted with the outlying entry in cell (1,1) eliminated, as if it were a missing value

		Second Factor			Row Mean	Row Effect, R_i
		1	**2**	**3**		
First Factor	**1**		11	9	10	3
	2	12	7	5	8	1
	3	8	3	1	4	–3
	Column Mean	10	7	5	**Overall Mean** = 7	
	Column Effect, C_j	3	0	–2		**Overall Effect, E** = 7

Table 8. Mean-based fitted values f_{ij} from Table 7

f_{ij}		Second Factor		
		1	**2**	**3**
First Factor	**1**		10	8
	2	11	8	6
	3	7	4	2

Table 9. Mean-based residual values r_{ij} from Tables 7 and 8

r_{ij}		Second Factor		
		1	**2**	**3**
First Factor	**1**		1	1
	2	1	–1	–1
	3	1	–1	–1

6. Method 3: Median polish

For the data in Table 3, median polish stops in two steps, when starting either with rows or columns, to yield Table 11. See Tables 10 and 11 for the two steps of median polish. The fitted values are in Table 12. We have continued to use the words *overall effect, row effect*, and *column effect*, but they are medians or from a procedure using medians, unlike in all of the other sections, where they indicate means. *Minitab*™ contains a dropdown tab for performing median polish, where the user can determine whether row or columns are polished first and the number of steps to be performed.

Subtracting the analogous entries in Tables 3 and 12, we see that all residual values are zero except for cell $(1,1)$, where the residual value is $r_{11} = 34 - 16 = 18$, which is shown in Table 11, as well. This method has given the outlying component 18 and left a fitted value of 16 in cell $(1,1)$, in Table 12. This accomplishes the twin goals of fitting a table and isolating the troublesome portion of the entry in cell $(1,1)$.

Although median polish is a useful exploratory data analysis tool and works very well for this example, there is a major problem. If half or more of a row were outliers (for example, two entries in row 1 of a 3×3 table), the median of that row would be compromised and the method would breakdown without giving the user a warning. In other words, the polishing would be performed uninterrupted, but the results might be meaningless or misleading [4, 5].

By comparing Tables 2 and 12, we see that there are two different additive fits of Table 1. In Table 2, the overall effect is 8, the row effects are 4, 0, and –4, and the column effects are 4, –1, and –3. In Table 12, the overall effect is 7, the row effects are 4, 0, and –4, and the column effects are 5, 0, and –3, but the fitted values are the same. Two additive fits are possible, depending upon the criteria. In Table 2, the mean of the row effects and the mean of the column effects are zero. In Table 12, the median of the row effects and the median of the column effects are zero.

Table 10. First step of median polish of Table 3, beginning with rows

			Row Median	**Row**	**Medians**	**Subtracted**
34	11	9	11	23	0	–2
12	7	5	7	5	0	–2
8	3	1	3	5	0	–2

Table 11. Second polishing step, which is by columns of the result in Table 10

	23	0	–2	11
	5	0	–2	7
	5	0	–2	3
Column Median	5	0	–2	7
Column	18	0	0	4
Medians	0	0	0	0
Subtracted	0	0	0	–4

Table 12. Fitted values as a result of median polish in Tables 10 and 11

f_{ij}		**Second Factor**			**Row**
		1	**2**	**3**	**Effect**
First Factor	**1**	16	11	9	4
	2	12	7	5	0
	3	8	3	1	–4
Column Effect		5	0	–2	**Overall Effect** = 7

7. Method 4: The new recommended method

This method temporarily replaces the entry in each cell (i, j) containing an outlier with the placeholder t_{ij}, which equals a numerical value resulting in a residual of zero for the cell in a mean-based, that is, least-squares fit. When there is more than one outlier, all the placeholders' values are determined simultaneously by a system of linear equations. A table with two outliers is analyzed in Section 8.

Referring to Table 13 and setting the fitted value $f_{11} = E + R_1 + C_1$ equal to $t_{11} = t$ gives

$$\frac{t + 56}{9} + \frac{2t + 4}{9} + \frac{2t + 4}{9} = t$$

which has the unique solution $t = 16$. Substituting this value for t into Table 13, we see that all residuals are zero, except $r_{11} = 34 - 16 = 18$. The observation 34 in cell (1,1) has been additively decomposed into the outlying component 18, which can be set aside for separate examination, and

a remaining component 16, which can be used in subsequent analyses of the table. In those analyses, the residual will be zero for cell $(1,1)$, when replacing the original entry with $t = 16$. This has fulfilled our goal for investigating the table.

This method can be applied to tables that have all but one well-chosen entry in any row or column deemed to be outliers [3]. In particular, for an $r \times c$ table, the whole $r-1$ by $c-1$ lower-right block of entries can be outliers or otherwise considered contaminated, yet this method can successfully proceed by replacing each entry in the block with its own placeholder. Of course, that many compromised entries would threaten the integrity of the experiment.

This method has a fail-safe protection, because, if the method breaks down for any reason, such as having a whole row of outliers, the set of equations for the placeholders has no solution. In those cases, the determinant of the set of linear equations for the placeholder has a determinant zero [3].

Table 13. Mean-based fitting of Table 3 with the placeholder $t_{11} = t$ substituted for the outlying entry in $(1,1)$

		Second Factor			**Row**	**Row**
		1	**2**	**3**	**Mean**	**Effect, R_i**
First Factor	**1**	t	11	9	$(t+20)/3$	$(2t+4)/9$
	2	12	7	5	8	$(16-t)/9$
	3	8	3	1	4	$(-20-t)/9$
	Column Mean	$(t+20)/3$	7	5	**Overall Mean =** $(t+56)/9$	
	Column Effect, C_j	$(2t+4)/9$	$(7-t)/9$	$(-11-t)/9$		**Overall Effect, E =** $(t+56)/9$

8. Method 4: Another example

Consider the observations in Table 14. The diagnostic graph in Figure 2 shows that there are outliers in cells $(1,1)$and $(2,3)$ and that the observations appear to be sufficiently planar otherwise. Although the entries in the two suspect cells are the same size, they have been visually separated by rotating the axes.

Replacing $x_{11} = 35$ with t_{11} and replacing $x_{23} = 35$ with t_{23} give

$$E = \frac{t_{11}+t_{23}+89}{12},\ R_1 = \frac{t_{11}+36}{4} - E,\ R_2 = \frac{t_{23}+31}{4} - E,\ C_1 = \frac{t_{11}+24}{3} - E, \text{ and } C_2 = \frac{t_{23}+14}{3} - E,$$

so that the fitted values are

$$f_{11} = E + R_1 + C_1 = t_{11} = \frac{6t_{11}-t_{23}+115}{12} \text{ and } f_{23} = E + R_2 + C_3 = t_{23} = \frac{-t_{11}+6t_{23}+60}{12},$$

or

$$6t_{11} + t_{23} = 115 \text{ and } t_{11} + 6t_{23} = 60.$$

The unique solution is $t_{11} = 18$ and $t_{23} = 7$. In Table 14, replacing $x_{11} = 35$ with $t_{11} = 18$ and replacing $x_{23} = 35$ with $t_{23} = 7$ produces a table that is ready for least-squares fitting or a statistical procedure such as ANOVA. The residuals of the least-squares fit will be zero in cells (1,1) and (2,3). The discordant portions of the problematical observations are $r_{11} = x_{11} - t_{11} = 35 - 18 = 17$ and $r_{23} = x_{23} - t_{23} = 35 - 7 = 28$.

Table 14. A table with two outliers, which is examined in Section 8

x_{ij}		. Second Factor .			
		1	**2**	**3**	**4**
First Factor	**1**	35	16	11	9
	2	14	12	35	5
	3	10	8	3	1

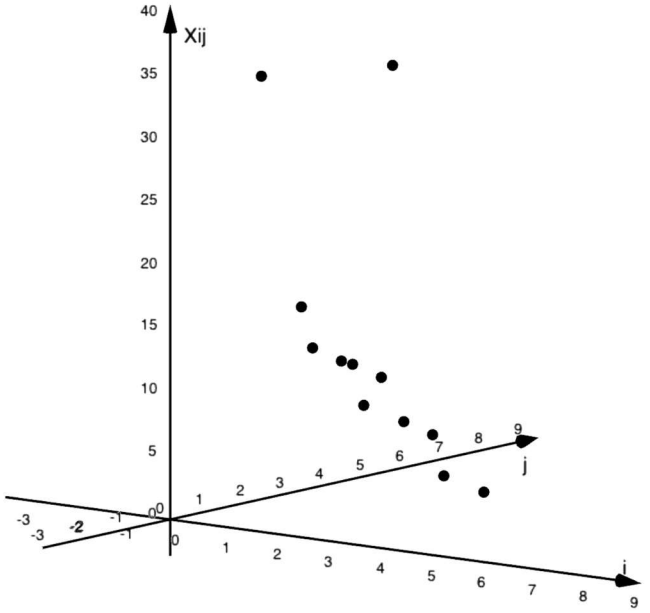

Figure 2. Diagnostic graph for Table 14, which shows two outliers and planarity Otherwise

9. Conclusions

A relatively new method for handling outliers in tables of measurement data when there are no replications has been presented and compared to other methods. This method is recommended for the classroom at all educational levels, as well as for statistical practice. We have found that students have no difficulty understanding the reasons that outliers are troublesome, but analyses such as the one in Section 4 can be useful to show them. Students appear to find it particularly enjoyable to examine data that have outliers and have not been recognized or mitigated, but nonetheless have been statistically analyzed in the literature or textbooks. Besides applying the method to tables of data, this topic is a source for projects and interesting homework. For example, students can be asked to implement the method, or parts of it, in the statistical computing package that they are using in class. The method's mathematics, which appears in [3], might be helpful for that. Students could consider or discuss other options for ways to assign values to the placeholders t_{ij}, such as using a different criterion or adding random noise. Students could be asked to examine non-public data sets to which they have access, such as in their science or psychology labs.

We have called the troublesome observations *outliers*. However, that designation is not required for the new method to proceed. The method can be applied to the entry in a cell that may be from an untrusted source or the measurement of that entry may be suspicious for a variety of reasons. The method can be used to replace the entry with a value that is inconsequential for the further analyses of the data set, and the portion that would be the discordant part may be ignored in those cases. Observe that the troublesome entry need not be far from the data in order to apply this method. The new method can be used for missing values, where their placeholders will have a value for subsequent analyses and the discordant part is simply its negation. This allows computations of tables with missing values without having the flaw that is displayed in Section 5.

References

[1]. Aggarwal, C. C. (2017). *Outlier Analysis* (2nd ed.). Berlin, Germany: Springer Nature. *doi.org/10.1007/978-3-319-47578-3*

[2]. American Society for Testing Materials Subcommittee E11.10 on Sampling/Statistics (2021). *ASTM E178-21: Standard Practice for Dealing with Outlying Observations.* West Conshohocken, PA, USA: ASTM International, *www.astm.org/e0178-21.html. doi.org/10.1520/E0178-21*

[3]. Farnsworth, D. L. (2023). Modeling and fitting two-way tables containing outliers. *International Journal of Mathematics and Mathematical Sciences*, 50 (February), 1–6, Article ID 6352058. *doi.org/10.1155/2023/6352058*

[4]. Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (1983). *Understanding Robust and Exploratory Data Analysis*. New York, NY, USA: Wiley.

[5]. Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (1985). *Exploring Data Tables, Trends, and Shapes*. New York, NY, USA: Wiley.

[6]. Little, R. J. A., and Rubin, D. B. (2020). *Statistical Analysis with Missing Data* (3rd ed.). New York, NY, USA: Wiley. *doi.org/10.1002/9781119482260*

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

[8]. Suri, N. N. R. R., Murty, M. N., and Athithan, G. (2019). *Outlier Detection: Techniques and Applications*. Berlin, Germany: Springer. *doi.org/ 10.1007/978-3-030-05127-3*
[9]. Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA, USA: Addison-Wesley.