A Research of How to Distinguish Suspicious Data

Sheng-zhi Guo [1, a], Li-sen Pu[1] and Cheng-wei Yang[1]

[1] BAICHENG ORDNANCE TEST CENTER OF CHINA，BAICHENG 137000，China

[a]552449556@qq.com

Keywords: Outlier Judge.

Abstract. We get a lot of data from Light weapons testing, such as the data of dispersion and intensive inspection. It is necessary to distinguish the suspicious data to ensure the result of the examination. This essay introduces some criterions about how to judge the outlier in small sample from normal populations, and give some examples to illustrate how to distinguish the data, which can be referred when doing data processing.

Introduction

The suspicious data often appear in Light weapons testing, and it will affect the accuracy of Firing table if judged as outliers or not. An outlier is a value which diverge away from other values in a set of observational results, furthermore, is not belong to the same population.

Suppose $X_1,X_2,\ldots,X_n$ are samples from normal population $N(\mu,\sigma^2)$, in which μ is unknown, σ^2 is known or not, we arrange $X_1,X_2,\ldots,X_n$ from small to large, as $X_{(1)}\le X_{(2)}\le\ldots\le X_{(n)}$.Here $X_{(n)}$(or $X_{(1)}$)diverge obviously away from others, but it can not be set as an outlier, it is called suspicious result instead, if it belongs to the same population with other $X_{(i)}(i\neq 1,n)$, it is not an outlier. Or, we call it an outlier, and it must be rejected.

Some major methods for determining outliers

1 A method for extremal deviation when σ is known, μ is unknown

Judge if there is an outlier in sample values $X_1,X_2,\ldots,X_n$ or not when the overall standard deviation σ is known.

(1) We arrange $X_1,X_2,\ldots,X_n$ from small to large, as $X_{(1)}\le X_{(2)}\le\ldots\le X_{(n)}$, we regard $X_{(n)}$(or $X_{(1)}$) as suspicious data.

(2) Calculate the average of sample values, $\overline{X}$.

(3) Calculate the statistics, G_n.

$$G_n = \frac{X_{(n)} - \overline{X}}{\sigma}\,(\text{or } G_n = \frac{\overline{\mathrm{X}} - \mathrm{X}_{(1)}}{\sigma}) \qquad (1)$$

(4)Significant level, α, is given, we can get β_α through check Table 1 with n and α to make $P\{\, G_n > \beta_\alpha\} = \alpha$.

(5) Judgment: If $G_n > \beta_\alpha$, then reject H_0.It means that $X_{(n)}$(or $X_{(1)}$)is an outlier, which should be rejected. But if $G_n \le \beta_\alpha$, we cannot reject H_0. It means that there is no reason to consider $X_{(n)}$(or $X_{(1)}$) as an outlier, so we must keep it.

Table 1: The critical value distribution of $G_n(\alpha=P\{\, G_n > \beta_\alpha\})$

n	10	11	12	13	14	15	16	17	18	19	20
$t_{0.01}$	2.931	2.973	3.010	3.043	3.071	3.099	3.214	3.147	3.168	3.188	3.207

2 A method for extremal deviation when σ and μ are unknown

Judge if there is an outlier in sample values $X_1,X_2,\ldots,X_n$ or not.

(1) We arrange $X_1,X_2,\ldots,X_n$ from small to large, as $X_{(1)}\le X_{(2)}\le\ldots\le X_{(n)}$, we regard $X_{(n)}$(or $X_{(1)}$)as suspicious data.

(2) Calculate the average and standard deviation of sample values, $\overline{X}$ and S.

(3) Calculate the statistics, *Q*.

$$Q = \frac{X_{(n)} - \overline{X}}{S} \quad \text{（or} \quad Q = \frac{\overline{\mathrm{X}} - \mathrm{X}_{(1)}}{S} \text{）} \qquad (2)$$

(4)Significant level, *α*, is given ,we can get q_α through check Table 2 with *n* and *α*.

(5) Judgment: If $Q>q_\alpha$, then reject H_0.It means that $X_{(n)}$(or $X_{(1)}$)is an outlier, which should be rejected. But if $Q\le q_\alpha$, we cannot reject H_0. It means that there is no reason to consider $X_{(n)}$(or $X_{(1)}$)as an outlier, so we must keep it.

Table 2: The critical value distribution of $Q(\alpha=P\{ Q>q_\alpha\})$

n	10	11	12	13	14	15	16	17	18	19	20
$t_{0.05}$	3.176	2.234	2.285	2.331	2.371	2.408	2.443	2.475	2.504	2.527	2.557

3 A method for the ratio of range

Judge if there is an outlier in sample values $X_1,X_2,\dots,X_n$ or not.

(1) We arrange $X_1,X_2,\dots,X_n$ from small to large, as $X_{(1)}\le X_{(2)}\le\dots\le X_{(n)}$, we regard $X_{(n)}$(or $X_{(1)}$)as suspicious data.

(2) Calculate the statistics, r_{ij}.

There are four different algorithms for this statistics:

$$r_{10} = \frac{X_{(n)} - X_{(n-1)}}{X_{(n)} - X_{(1)}} \quad \text{（or} \quad r_{10} = \frac{X_{(2)} - X_{(1)}}{X_{(n)} - X_{(1)}} \text{ ）} \qquad (3)$$

$$r_{11} = \frac{X_{(n)} - X_{(n-1)}}{X_{(n)} - X_{(2)}} \quad \text{（or} \quad r_{11} = \frac{X_{(2)} - X_{(1)}}{X_{(n-1)} - X_{(1)}} \text{ ）} \qquad (4)$$

$$r_{21} = \frac{X_{(n)} - X_{(n-2)}}{X_{(n)} - X_{(2)}} \quad \text{（or} \quad r_{21} = \frac{X_{(3)} - X_{(1)}}{X_{(n-1)} - X_{(1)}} \text{ ）} \qquad (5)$$

$$r_{22} = \frac{X_{(n)} - X_{(n-2)}}{X_{(n)} - X_{(3)}} \quad \text{（or} \quad r_{22} = \frac{X_{(3)} - X_{(1)}}{X_{(n-2)} - X_{(1)}} \text{ ）} \qquad (6)$$

The statistics, r_{ij}, is depend on *n*.According to the results of stochastic simulation, we usually think like this: When 3≤n≤7, r_{10} is the best one.When 8≤n≤10, r_{11} is the best one. When 11≤n≤13 , r_{21} is the best one. When 14≤n≤30 , r_{22} is the best one.

(3)Significant level, *α*, is given, we can get the critical value, r_α , from look-up the Table 3 of the critical value distribution of r_{ij} to make $P\{ r_{ij} >r_\alpha\}=\alpha$.

(4) Judgment: If $r_{ij} >r_\alpha$, then reject H_0.It means that $X_{(n)}$(or $X_{(1)}$) is an outlier, which should be rejected. But if $r_{ij}\le r_\alpha$, we cannot reject H_0. It means that there is no reason to consider $X_{(n)}$(or $X_{(1)}$) as an outlier, so we must keep it.

Table 3: The critical value distribution of R_{ij}

n	5	6	7	8	9	10	11	12	13	14	15	16
$t_{0.05}$	R_{10}			R_{11}			R_{21}			R_{22}		
	0.642	0.560	0.507	0.554	0.512	0.477	0.576	0.546	0.521	0.546	0.525	0.507

4 A method for judging suspicious coordinate values

Judge if there is an outlier in sample values (X_1,Y_1), (X_2,Y_2),…, (X_n,Y_n) or not.

(1) Pick the most different sample, Z_α, then Z_α is suspicious coordinate.

(2) Calculate the average and intermediate error of sample values without Z_α.

$$\overline{Z}_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n-1} Z_i \qquad (7)$$

$$E_{n-1} = 0.6745\sqrt{\frac{\sum_{i=1}^{n-1}\left(Z_i - \overline{Z}_{n-1}\right)^2}{n-2}} \tag{8}$$

(3) Calculate the statistics, *t*.

$$t = \frac{\left|Z_\alpha - \overline{Z}_{n-1}\right|}{E_{n-1}} \tag{9}$$

(4) We can get the value of t_α with the number of shooting when confidence coefficient, α, equal to 0.01.

(5) Judgment: If $t > t_\alpha$, then Z_α is an abnormal coordinate value,which should be rejected. But if $t \le t_\alpha$, then Z_α is a normal coordinate value, we must keep it.

Table 4: The value of t_α when α equal to 0.01

n	10	11	12	13	14	15	16	17	18	19	20
$t_{0.01}$	7.03	6.79	6.60	6.45	6.34	6.24	6.17	6.10	6.05	6.00	5.97

Application examples

There are a set of coordinates which a weapon’s bullet hole in 200 meter. Judge if there is an outlier or not.

Table 5

（-48.8,78.7）	（-96.0,98.0）	（-57.5,100.0）	（-53.3,104.9）	（-55.7,109.0）
（-66.7,111.7）	（-49.8,112.4）	（-50.5,116.6）	（-38.0,117.0）	（-59.8,118.6）
（-52.0,120.0）	（-54.2,119.8）	（-60.9,120.3）	（-55.2,121.4）	（-46.4,119.5）
（-67.9,133.0）	（-41.1,137.9）	（-40.5,144.1）	（-54.6,133.0）	（-58.5,118.4）

(1) Pick the most different sample, Z_α.

Calculate the average of these 20 coordinates, we can get (-55.4, 116.7) is the answer.

Pick the farthest coordinate from average coordinate, we can know the farthest coordinate is (-96.0, 98.0).

(2) Calculate the average and intermediate error of sample values without (-96.0, 98.0), abscissa and ordinate was calculated respectively. The average and intermediate error of abscissa is: $\overline{Z}_{19}$=-53.2, E_{19}=5.47. The average and intermediate error of ordinate is: $\overline{Z}_{19}$=117.7, E_{19}=9.73.

(3) Calculate the statistics, *t*.

$$t_{abscissa} = \frac{\left|-96.0-(-53.2)\right|}{5.47} \approx 7.82 \tag{10}$$

$$t_{ordinate} = \frac{\left|98.0-117.7\right|}{9.73} \approx 2.52 \tag{11}$$

(4) t_α equal to 5.97 when shooting 20 times and the confidence coefficient, α, equal to 0.01

(5) Judgment: $t_{abscissa} > t_\alpha$, $t_{\text{ordinate}} < t_\alpha$, so abscissa is an abnormal coordinate value, which should be rejected, therefore , this bullet hole’s coordinate shouldn’t be used.

Summary

Outlier always have great influence on test results.If we cannot judge it accurately and eliminate it immediately, it will cause a serious consequence. This essay introduces some methods about how to judge the outlier in small sample from normal populations, and give some examples to illustrate how to distinguish the data. These methods can be used in data processing as a criterion.

References

[1] Zhang-geng Yan: *Technology Of Make Firing Table* (National Defense Industry Press, Beijing 2002).

[2] De-bao Li and Shu-sen Guo:*The Methods Of Conventional Weapon's Stereotype- Make Firing Table For Firearms* (Defense Technology Committee ,Beijing 1988).

[3] Shu-yuan He: *Probability theory and mathematical statistics*(Higher Education Press, Beijing 2008).

[4] Ke-hui Dai,Song-yuan Fang and De-xiang Xia in: *Statistical found of abnormal data* in Application of Statistics and Management , edited by Chinese Association for Applied Statistics, Beijing (1984).

[5]Zhi-zhong Guo: *Higher Mathematics* (Tsinghua University Press, Beijing 2012).