Maximum Z Scores and Outliers

Ronald E. Shiffler


Stable URL:
http://links.jstor.org/sici?sici=0003-1305%28198802%2942%3A1%3C79%3AMZSAO%3E2.0.CO%3B2-W

*The American Statistician* is currently published by American Statistical Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR’s Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR’s Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/astata.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.
readily obtain
\[ \int_0^c \frac{2^{n-1}}{\pi} \int_{-\infty}^{\infty} \frac{e^{r2}}{(1 + e^{r2})^n} \frac{\sin(\alpha t)}{t} \, dt, \quad (A.1) \]
and, upon introducing the integration variable \( z = e^{\alpha t} \), in (A.1) we obtain
\[ \Pr(\bar{X} \leq c) = .5 + \frac{2^{n-1}}{\pi} \int_0^{\infty} \frac{z^{(n-2)/2} \sin(nc \log z)}{(1 + z)^n \log z} \, dz, \]
where, without loss of generality, we have chosen \( c > 0 \).

[Received October 1986, Revised April 1987.]

RONALD E. SHIFFLER*

The magnitude of \( Z(n) \), the Z score associated with the largest value of \( X \) in a data set of size \( n \), is shown to be bounded above by \((n - 1)/\sqrt{n}\). As a result, outliers defined as values exceeding four standard deviations from the mean cannot exist for small data sets.

KEY WORDS: Largest Z score; Upper bound.

1. INTRODUCTION

How large can the largest Z score be for a set of raw data? Huck, Cross, and Clark (1986) argued that \( Z(n) \), the Z score corresponding to the largest value of \( X \), cannot exceed \((n - 1)^{1/2}\). Shiffler (1987) showed that \( Z(n) \) is strictly less than \((n - 1)^{1/2}\) and gave an achievable lower bound of \((1/n)^{1/2}\). This note gives an achievable upper bound for \( Z(n) \) and shows that outliers, defined in terms of standard deviations from the mean, cannot exist for small data sets.

2. DEVELOPMENT

Let \( X_1, X_2, \ldots, X_{n-1} \) be a random sample of size \( n - 1 \) from a population with unknown mean and variance and, without loss of generality, assume the \( X_i \) are already ordered and \( \bar{X}_{n-1} = 0 \). Suppose another observation \( X_n \) is added to the sample such that \( X_n > X_{n-1} \). The mean of the \( n \) values is \( \bar{X}_n/n \), and the variance estimate, \( S^2_n = \Sigma(X_i - \bar{X}_n)^2/(n - 1) \), can be written as
\[ S^2_n = [(n - 2)/(n - 1)] \cdot S^2_{n-1} + X^2_n/n. \quad (1) \]
The largest positive Z score for the \( n \) values,
\[ Z(n) = (X_n - X_{n}/n)/S_n, \quad (2) \]
is maximized whenever \( S_n \) is minimized. From (1) this occurs when \( S^2_{n-1} = 0 \); hence, Equation (2) reduces to
\[ Z(n) = (n - 1)/\sqrt{n}. \quad (3) \]
Equation (3) is the maximum achievable value for the largest positive Z score based on \( n \) values. Table 1 lists this maximum for several sample sizes. The smallest achievable value for the largest negative Z score is \(-(n - 1)/\sqrt{n}\).

3. IMPLICATIONS

Outliers are sometimes defined as those values for which \( |Z| \) exceeds 3 (Sincich 1986) or 4 (Younger 1979). As Table 1 indicates, definitions such as these preclude the existence of outliers in samples of size \( n \leq 10 \) for three standard deviations or \( n \leq 17 \) for four standard deviations. For example, the data set \( \{0, 0, 0, 0, 1\} \) yields \( Z(5) = 1.789 \). Thus, by the three or four standard deviation definition, the value of 1 million is not an outlier. Alternatively, a data set with 17 0s and one 1 produces \( Z(18) = 4.007 \), exposing the value 1 as an outlier.

Tukey (1977) defined outliers as values beyond the outer fences. The values of 1 million in the \( n = 5 \) data set and 1 in the \( n = 18 \) data set both qualify as outliers; however, Tukey’s definition also has its drawbacks. No matter how small the 18th value in the data set \( \{0, 0, \ldots, 0, X_{18}\} \), it will be an outlier according to both Tukey and the four standard deviation definition.

<table>
<thead>
<tr>
<th>( n )</th>
<th>( Z(n) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1.155</td>
</tr>
<tr>
<td>4</td>
<td>1.500</td>
</tr>
<tr>
<td>5</td>
<td>1.789</td>
</tr>
<tr>
<td>10</td>
<td>2.846</td>
</tr>
<tr>
<td>11</td>
<td>3.015</td>
</tr>
<tr>
<td>17</td>
<td>3.881</td>
</tr>
<tr>
<td>18</td>
<td>4.007</td>
</tr>
</tbody>
</table>

**Table 1. Maximum Absolute Z Score per Sample Size**

*Ronald E. Shiffler is Associate Professor, School of Business, University of Louisville, Kentucky 40292. The author thanks the editors for several suggestions that improved the clarity of the article.

© 1988 American Statistical Association

The American Statistician, February 1988, Vol. 42, No. 1
4. CONCLUSION

The concept of a Z score as a measure of a value’s position within a data set in terms of standard deviations is intuitively appealing. Unfortunately, the behavior of \( Z_{(n)} \) is quite constrained for small data sets of size \( n \). When \( n \leq 17 \), \( |Z_{(n)}| \) cannot exceed 4 regardless of the configuration of values. Consequently, no value can be an outlier according to a four standard deviation definition.

[Received April 1987. Revised June 1987.]

REFERENCES


---

Computer Programs to Demonstrate Some Hypothesis-Testing Issues

MAX E. JERRELL*

Three computer programs are presented, which can be used to demonstrate some issues that should be considered in conducting hypothesis tests.

KEY WORD: Costs.

1. INTRODUCTION

Students in introductory statistics courses often have difficulty understanding some of the factors that should be considered in hypothesis testing. This paper presents some MINITAB programs (see Ryan, Joiner, and Ryan 1985) that demonstrate (a) how the costs of Type I and Type II errors affect the choice of a rejection region, (b) how increased sample size reduces the probable cost of an error, and (c) the effect of sampling costs on the hypothesis-test procedure. These programs have been kept simple so that students can experiment with them.

The programs consider determining the “best” rejection region for the simple hypothesis

\[
H_0: \mu = 10, \quad H_A: \mu = 20.
\]

The best rejection region is assumed to be the one that minimizes the cost function

\[
\text{cost} = \alpha C_1 + \beta C_\Pi,
\]

where \( C_1 \) and \( C_\Pi \) are the costs of Type I and Type II errors. It is assumed that the samples are taken from a normal distribution.

2. DESCRIPTION OF PROGRAMS

Program 1 demonstrates the effect of the cost of errors on the rejection region. The program’s output is shown in Figure 1. In this graph the points labeled A indicate the value of the cost function with \( C_1 = C_\Pi = 10,000 \) and the points labeled B are for the cost function with \( C_1 = 10,000 \) and \( C_\Pi = 5,000 \). The program calculates the cost function using points to the right of the abscissa as the rejection region. The best rejection region is assumed to be the one that minimizes the cost function. It can be seen that the costs of Type I and Type II errors are important considerations in selecting rejection regions.

Program 2 demonstrates the effect of an increased sample size on the problem. Its output is shown in Figure 2, where A indicates the values of the cost function for a sample of size 30 and B indicates the values for a sample of size 300. This output indicates increased sample size and reduces the value of the cost function in the region near the best rejection region, but it does not do so for less fortunate regions.

Program 3 includes sampling costs. Here it is assumed that sampling costs are a linear function of the sample size and are included by adding them to the cost function. The output is shown in Figure 3 with NCOST the sum of the cost function and sampling costs and NOBS the sample size.

---

*Max E. Jerrell is Assistant Professor, Department of Economics, Northern Arizona University, Flagstaff, Arizona 86001. The author wishes to thank the associate editor and referees for their suggestions to improve the original draft.

Figure 1. Values of the Cost Function Calculated for Various Critical Values. The rejection region is to the right of the abscissa. Points labeled A are for the cost combination \( C_1 = C_\Pi = 10,000 \); those labeled B are for the cost combination \( C_1 = 10,000 \) and \( C_\Pi = 5,000 \).