Trimmed Estimators of the Average and Standard Deviation

Dr. Wayne A. Taylor

Abstract: The Distribution Analyzer software package identifies points that are potential outliers by reporting all values more than 4.5 standard deviations from the average. It uses robust estimators for the average and standard deviation to avoid potential outliers from biasing the estimates. Distribution Analyzer uses the trimmed estimators described in this article for this purpose.

1.0 Introduction

A trimmed estimator is obtained by eliminating T% of the extreme values before calculating the estimate. This results in a more robust statistic when outliers are present. An example is the median, trimming all but the middle 1 or 2 values.

A procedure is provided for calculating 20% trimmed estimates of the average and standard deviation. In the case of the average, taking the average of the trimmed values still results in an unbiased estimate of the true average. In the case of the standard deviation, estimating the standard deviation from the trimmed deviations results in underestimating the standard deviation. Table 2 contains correction factors for the trimmed deviation estimates to obtain unbiased estimates of the true standard deviation.

2.0 Calculating the Trimmed Average Estimator

For a sample size of n, let $X_1, X_2, \ldots, X_n$ represent the data. Sort the data into order from the smallest to the largest, represented by $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. The values used to calculate the T% trimmed estimator are determined as follows:

$$I_{mid} = \frac{n+1}{2}, \qquad I_{mid} \text{ is not always an integer value (n=4, } I_{mid} \text{=2.5).}$$

$$\text{Range} = \frac{n\left(1 - T/100\right)}{2}$$

$$I_{min} = \text{ceil}\left(I_{Mid} - \text{Range}\right) \text{ and } I_{max} = \text{floor}\left(I_{Mid} + \text{Range}\right)$$

where ceil() means to round up to the nearest integer and floor() means to round down to the nearest integer. Approximately T/2% are trimmed from each end. The trimmed estimate of the average is:

$$\bar{X}_{trimmed} = \frac{X_{(I_{min})} + \ldots + X_{(I_{max})}}{I_{max} - I_{min} + 1} \text{ with } n_{excluded} = n - \left(I_{max} - I_{min} + 1\right) \text{ and Actual } T\% = \frac{n_{excluded}}{n}$$

It is possible for an even sample size and T% close to 100% that $I_{min} > I_{max}$. In this case, reduce I_{min} by one and increase I_{max} by one, resulting in the median.

Table 1 shows the results for 20% trimming. 10% are trimmed from each end. No trimming occurs for 2-5 samples. One value is trimmed from each end for 6-15 samples. Two or more are trimmed from each end for 16 or more samples. Since estimates of the standard deviation are imprecise for small sample sizes, this is an adequate amount of trimming for most purposes. Trimming too many values uses less of the data, resulting in less accurate estimates.

Table 1: 20% Trimming

n	I_{mid}	Range	I_{min}	I_{max}	$n_{excluded}$	Actual T%	J_{max}
2	1.5	0.8	1	2	0	0%	2
3	2	1.2	1	3	0	0%	3
4	2.5	1.6	1	4	0	0%	4
5	3	2	1	5	0	0%	5
6	3.5	2.4	2	5	2	33%	4
7	4	2.8	2	6	2	29%	5
8	4.5	3.2	2	7	2	25%	6
9	5	3.6	2	8	2	22%	7
10	5.5	4	2	9	2	20%	8
11	6	4.4	2	10	2	18%	9
12	6.5	4.8	2	11	2	17%	10
13	7	5.2	2	12	2	15%	11
14	7.5	5.6	2	13	2	14%	12
15	8	6	2	14	2	13%	13
16	8.5	6.4	3	14	4	25%	12
17	9	6.8	3	15	4	24%	13
18	9.5	7.2	3	16	4	22%	14
19	10	7.6	3	17	4	21%	15
20	10.5	8	3	18	4	20%	16
21	11	8.4	3	19	4	19%	17
22	11.5	8.8	3	20	4	18%	18
23	12	9.2	3	21	4	17%	19
24	12.5	9.6	3	22	4	17%	20
25	13	10	3	23	4	16%	21
26	13.5	10.4	4	23	6	23%	20
27	14	10.8	4	24	6	22%	21
28	14.5	11.2	4	25	6	21%	22
29	15	11.6	4	26	6	21%	23
30	15.5	12	4	27	6	20%	24

3.0 Calculating the Trimmed Standard Deviation Estimator

The trimmed average estimator is used to calculate the absolute deviations from the average:

$$D_i = \left| X_i - \bar{X}_{trimmed} \right|$$

Sort the absolute deviations into order from the smallest to the largest represented by $D_{(1)}, D_{(2)}, \ldots, D_{(n)}$.

In this case T% is all trimmed from the top. Use $J_{max} = n - n_{excluded}$. Calculate the T% trimmed standard deviation estimate:

$$S_{trimmed} = F_{correction} \times \bar{D}_{trimmed} \quad \text{where} \quad \bar{D}_{trimmed} = \frac{D_{(1)} + \ldots + D_{(J_{max})}}{J_{max}}$$

$F_{correction}$ is required to correct for bias in the estimate because trimming the larger deviations results in a biased estimator.

The correction factor depends on n and T. For T=20%, Figure 1 shows a plot of the correction factors for sample sizes up to 100. The stair-step pattern results from increases in the number excluded per Table 1. As the sample size increases, the steps get smaller and converge to 1.7903.

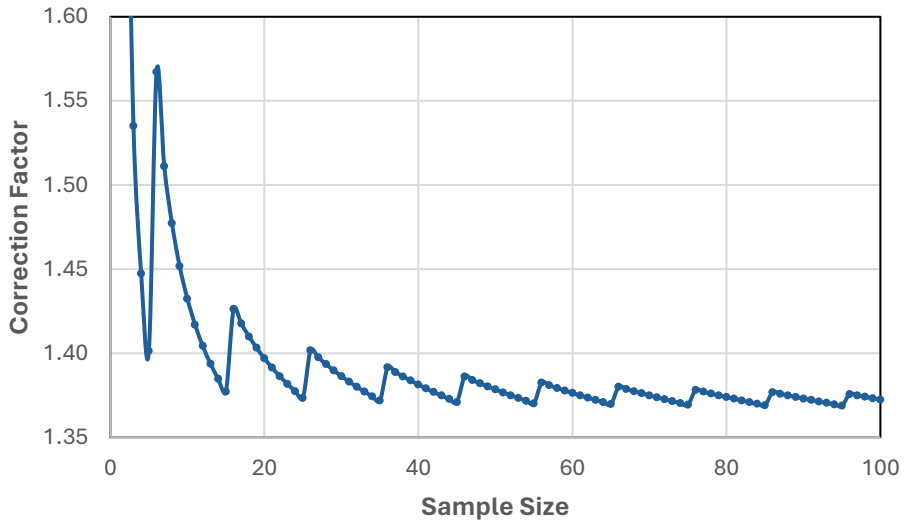

Figure 1: Correction Factors for T=20%

For T=20%, Table 2 gives correction factors for sample sizes up to 200. For larger sample sizes, the correction factor 1.7903 can be used. Each correction factor was obtained by simulation using 100 million normally distributed data sets, resulting in 4-digit accuracy.

Table 2: Correction Factors for 20% Trimming

n	Correction	n	Correction	n	Correction	n	Correction	n	Correction
>200	1.7903	41	1.7859	81	1.7877	121	1.7885	161	1.7889
2	1.7725	42	1.7723	82	1.7807	122	1.7838	162	1.7854
3	1.5351	43	1.7595	83	1.7739	123	1.7792	163	1.7818
4	1.4472	44	1.7473	84	1.7673	124	1.7746	164	1.7784
5	1.4012	45	1.7358	85	1.7609	125	1.7701	165	1.7750
6	2.3368	46	1.8502	86	1.8218	126	1.8117	166	1.8065
7	2.1388	47	1.8362	87	1.8147	127	1.8069	167	1.8029
8	2.0056	48	1.8229	88	1.8077	128	1.8022	168	1.7993
9	1.9095	49	1.8103	89	1.8010	129	1.7976	169	1.7958
10	1.8366	50	1.7981	90	1.7944	130	1.7931	170	1.7924
11	1.7794	51	1.7866	91	1.7880	131	1.7886	171	1.7890
12	1.7331	52	1.7756	92	1.7817	132	1.7843	172	1.7856
13	1.6948	53	1.7651	93	1.7757	133	1.7800	173	1.7823
14	1.6627	54	1.7551	94	1.7697	134	1.7757	174	1.7790
15	1.6351	55	1.7454	95	1.7639	135	1.7716	175	1.7758
16	1.9703	56	1.8392	96	1.8185	136	1.8101	176	1.8055
17	1.9218	57	1.8279	97	1.8121	137	1.8057	177	1.8021
18	1.8798	58	1.8171	98	1.8059	138	1.8013	178	1.7988
19	1.8434	59	1.8067	99	1.7999	139	1.7971	179	1.7955
20	1.8111	60	1.7967	100	1.7940	140	1.7929	180	1.7922
21	1.7826	61	1.7871	101	1.7882	141	1.7887	181	1.7890
22	1.7570	62	1.7778	102	1.7826	142	1.7847	182	1.7859
23	1.7339	63	1.7690	103	1.7770	143	1.7807	183	1.7828
24	1.7131	64	1.7604	104	1.7716	144	1.7767	184	1.7828
25	1.6941	65	1.7521	105	1.7664	145	1.7729	185	1.7797
26	1.8980	66	1.8316	106	1.8158	146	1.8087	186	1.7766
27	1.8713	67	1.8222	107	1.8100	147	1.8046	187	1.8047
28	1.8468	68	1.8130	108	1.8045	148	1.8006	188	1.8015
29	1.8243	69	1.8042	109	1.7990	149	1.7966	189	1.7952
30	1.8036	70	1.7957	110	1.7936	150	1.7927	190	1.7921
31	1.7845	71	1.7874	111	1.7884	151	1.7888	191	1.7891
32	1.7668	72	1.7795	112	1.7832	152	1.7850	192	1.7861
33	1.7503	73	1.7718	113	1.7782	153	1.7813	193	1.7831
34	1.7350	74	1.7643	114	1.7733	154	1.7776	194	1.7802
35	1.7206	75	1.7571	115	1.7684	155	1.7740	195	1.7773
36	1.8672	76	1.8260	116	1.8135	156	1.8075	196	1.8040
37	1.8488	77	1.8179	117	1.8083	157	1.8037	197	1.8009
38	1.8316	78	1.8100	118	1.8032	158	1.7999	198	1.7979
39	1.8154	79	1.8024	119	1.7982	159	1.7962	199	1.7950
40	1.8001	80	1.7950	120	1.7933	160	1.7925	200	1.7920

4.0 Effect of Trimming

Tables 3-5 show the effect of trimming on the accuracy of the estimates for different sample sizes. The values were obtained by simulating one million data sets from the standard normal distribution. In all cases, the average of the trimmed average estimates (Average column) is zero, showing that the trimmed average estimates are unbiased.

In Table 3 for n=20, the standard error of the trimmed average estimates (SE Ave) is 0.224 with 0% trimming and increases to 0.271 with 90% trimming (median). This is a 21% increase in the variation of the estimates of the average due to not using all the data.

The trimmed average deviation (Ave Dev) is less than 1 because no correction factor was applied. The relative standard deviation (RSD) of the trimmed average deviations (RSD AD) adjusts for differences in the trimmed average deviations (Ave Dev). The RSD increases from 17.3% to 95.7%. Trimming has a much greater effect on the standard deviation than the average.

Table 3: Effect of Trimming – Sample Size of 20

T%	I_{min}	I_{max}	J_{max}	Average	SE-Ave	Ave Dev	RSD-AD
0	1	20	20	0.000	0.224	0.778	17.3%
10	2	19	18	0.000	0.226	0.649	19.4%
20	3	18	16	0.000	0.230	0.552	21.6%
30	4	17	14	0.000	0.234	0.468	24.1%
40	5	16	12	0.000	0.238	0.392	27.1%
50	6	15	10	0.000	0.244	0.321	30.7%
60	7	14	8	0.000	0.249	0.254	35.7%
70	8	13	6	0.000	0.255	0.188	42.8%
80	9	12	4	0.000	0.262	0.125	54.7%
90	10	11	2	0.000	0.271	0.062	95.7%

Table 4: Effect of Trimming – Sample Size of 50

T%	I_{min}	I_{max}	J_{max}	Average	SE-Ave	Ave Dev	RSD-AD
0	1	50	50	0.000	0.142	0.790	10.8%
10	3	48	46	0.000	0.143	0.676	12.0%
20	6	45	40	0.000	0.146	0.556	13.7%
30	8	43	36	0.000	0.148	0.488	14.9%
40	11	40	30	0.000	0.152	0.395	17.1%
50	13	38	26	0.000	0.154	0.337	18.9%
60	16	35	20	0.000	0.158	0.255	22.5%
70	18	33	16	0.000	0.162	0.203	25.9%
80	21	30	10	0.000	0.167	0.126	34.2%
90	23	28	6	0.000	0.171	0.075	45.8%
100	25	26	2	0.000	0.175	0.025	98.0%

Table 5: Effect of Trimming – Sample Size of 100

T%	I_{min}	I_{max}	J_{max}	Average	SE-Ave	Ave Dev	RSD-AD
0	1	100	100	0.000	0.100	0.794	7.6%
10	6	95	90	0.000	0.102	0.656	8.6%
20	11	90	80	0.000	0.103	0.557	9.6%
30	16	85	70	0.000	0.105	0.473	10.7%
40	21	80	60	0.000	0.107	0.396	12.1%
50	26	75	50	0.000	0.109	0.324	13.7%
60	31	70	40	0.000	0.112	0.256	15.9%
70	36	65	30	0.000	0.115	0.190	19.0%
80	41	60	20	0.000	0.118	0.126	24.1%
90	46	55	10	0.000	0.121	0.063	35.5%
100	50	51	2	0.000	0.124	0.013	99.1%

Tables 4 and 5 show similar effects due to trimming for other sample sizes. 20% trimming is a compromise between removing enough points to eliminate outliers and not removing too many to reduce the accuracy of the estimates. Trimming affects the estimate of the standard deviation the most.

5.0 Conclusions

The trimmed estimates for the average and standard deviation resulted in a slight loss of accuracy for 20% trimming. For sample sizes of 16 or more, they remove at least two points from each end for the average and at least the four largest values for the standard deviation. For sample sizes of 26 or more, they remove at least three points from each end for the average and at least the six largest values for the standard deviation. This is sufficient for when there are multiple but a small number of outliers.

6.0 References

Distribution Analyzer 2.0, Dr. Wayne A. Taylor, Taylor Enterprises, Inc. (Variation.com/product/distribution-analyzer/).