Outlier Detection using Sequential Fences

Anwar Fitrianto
Department of Statistics
IPB University

IPB University
— Bogor Indonesia —

Outlier

- An observation that appears to be markedly deviated from other observations
- Great influence on the parametric data analyses and resulted in misleading results
- Initial screening of data is important before starting a data analysis (Tabachnick & Fidell, 2001).

Background

- Many traditional methods of detecting outliers are (only) good for symmetric distribution

- When the distribution of the data is skewed, then there are some mix-up

Related Terms

Swamping and Masking

- Swamping effect occurs when a second observation is labeled as an outlier in the presence of first outlier. After discarding the first outlying observation, the second observation is detected as clean observation.

- Masking can occur when we specify too few outliers in a test. For example, if we are testing for a single outlier when there are in fact two (or more) outliers

Outlier Detection by Schwertman et al. (2004)

He proposed a simple more general fences method which allows flexibility in setting the “outside rate”, that is, the probability that an observation from a non-contaminated normal population is outside a specified limit or boundary.

$$F_n = q_2 \pm \frac{\mathrm{IQR}}{k_n} Z_\alpha$$

Problem in Schwertman et al. (2004)

It assumes both a normal population and a large sample

Outlier Detection by Schwertman and de Silva (2007)

- ✓ Incorporating the sample size in the construction of fences
- ✓ a sequential procedure for identifying multiple outliers is suggested
- ✓ Plotting fences sequentially until there is no additional outlier is found
- ✓ It is less likely to misclassify an observation as an outlier

Sequential fences by Schwertman and de Silva (2007)

Fences for sample size *n* to identify *m*th outlier

$$F_{n,m} = q_2 \pm \frac{t_{df,\alpha_{nm}}}{k_n}(IQR)$$

- ✓ k_n and α_{nm} are values that are obtained from Table 1 and Table 2 in Schewertman and de Silva (2007) to construct m^{th} fences,
- ✓ $t_{df,\alpha_{nm}}$ is the value obtained from *t* distribution based on specified outside rate, α_{nm}
- ✓ df is calculated by equation below.

For $20 \leq n \leq 100$, the least squares quadratic equation for obtaining the degree of freedom, df, approaching *t* distribution based on the sample size is

$$df = 7.6809524 + 0.5294156n - 0.00237n^2.$$

Sequential fences by Schwertman and de Silva (2007)

- The sample sizes are adjusted to construct sequential fences using the Poisson model to decrease the tail probabilities, which are similar to that proposed by Davies and Gather (1993), and Gather and Becker (1997).
- By using Poisson model, the m contaminated observations can be checked. Let X be the number of outlying observations beyond the fence.

Based on the Poison model,

$$P(X < m) = e^{-n\alpha_{nm}}\left(1 + n\alpha_{nm} + \frac{(n\alpha_{nm})^2}{2!} + \cdots + \frac{(n\alpha_{nm})^{m-1}}{(m-1)!}\right) = 1 - \gamma.$$

Detection of Outlier Using Sequential Fences

- For example, to identify the first outlier, $m = 1$, with $\gamma = .05$, **means** there is a 0.95 probability of no outliers beyond the fence and only .05 probability that an observation beyond the fence is an outlier.

Sequential fences by Schwertman and de Silva (2007)

Table 1 from Schwertman and de Silva (2007). $IQR = k_n \sigma$

n	k_n	n	k_n	n	k_n	n	k_n	n	k_n	n	k_n
5	1.65798	22	1.33333	39	1.38071	56	1.34361	73	1.36635	90	1.34535
6	1.28351	23	1.4023	40	1.34165	57	1.3713	74	1.34454	91	1.36267
7	1.51475	24	1.33753	41	1.38021	58	1.34329	75	1.36557	92	1.34562
8	1.32505	25	1.40096	42	1.34104	59	1.37004	76	1.34495	93	1.36258
9	1.50427	26	1.33587	43	1.37779	60	1.34394	77	1.36543	94	1.3455
10	1.31212	27	1.39455	44	1.34226	61	1.36981	78	1.34478	95	1.3621
11	1.45768	28	1.33894	45	1.37737	62	1.34366	79	1.36474	96	1.34576
12	1.32968	29	1.39355	46	1.34175	63	1.36871	80	1.34514	97	1.36201
13	1.45268	30	1.3377	47	1.37536	64	1.34424	81	1.36461	98	1.34565
14	1.32353	31	1.38876	48	1.34278	65	1.36851	82	1.34499	99	1.36157
15	1.42975	32	1.34004	49	1.37501	66	1.34399	83	1.36398	100	1.34588
16	1.33318	33	1.38799	50	1.34235	67	1.36737	84	1.34532	200	1.3474
17	1.42684	34	1.33909	51	1.37331	68	1.3445	85	1.36387	300	1.34792
18	1.32959	35	1.38428	52	1.34322	69	1.36737	86	1.34517	400	1.348118
19	1.41322	36	1.34092	53	1.37301	70	1.34429	87	1.3633	∞	1.34898
20	1.33568	37	1.38367	54	1.34285	71	1.3665	88	1.34548		
21	1.41132	38	1.34017	55	1.37156	72	1.34474	89	1.36319		

Sequential fences by Schwertman and de Silva (2007)

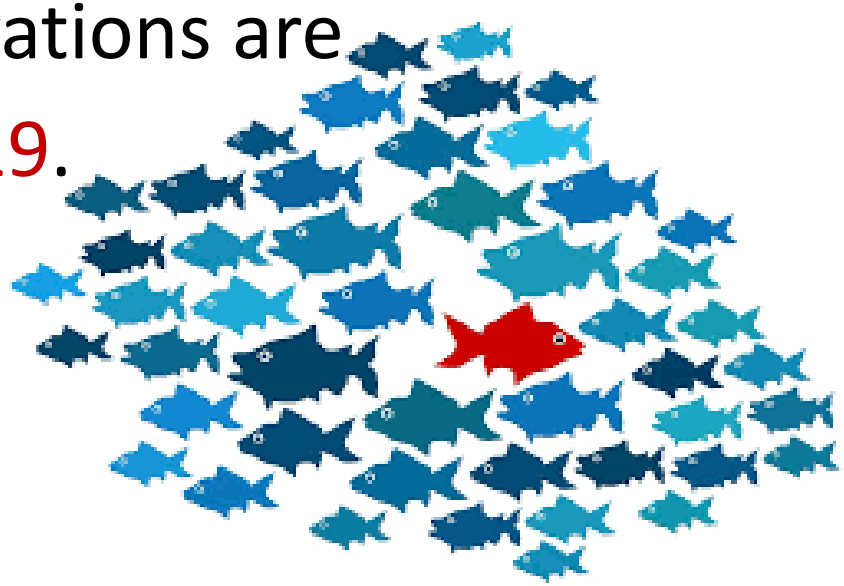
Table 2 from Schwertman and de Silva (2007): $C_m = n\ \alpha_{nm}$

1-γ	m=1	2	3	4	5	6
.75	.287682	.961279	1.72730	2.53532	3.36860	4.21920
.80	.223144	.824388	1.53504	2.29679	3.08954	3.90366
.90	.1053605	.531812	1.10207	1.74477	2.43259	3.15190
.95	.0512932	.355362	.817691	1.36632	1.97015	2.61301
.975	.025318	.242209	.618672	1.08987	1.62349	2.20189
.99	.0100503	.148555	.436045	.823249	1.27911	1.78528
.995	.005013	.103495	.337873	.672207	1.07793	1.53691

Example Wood specific gravity data

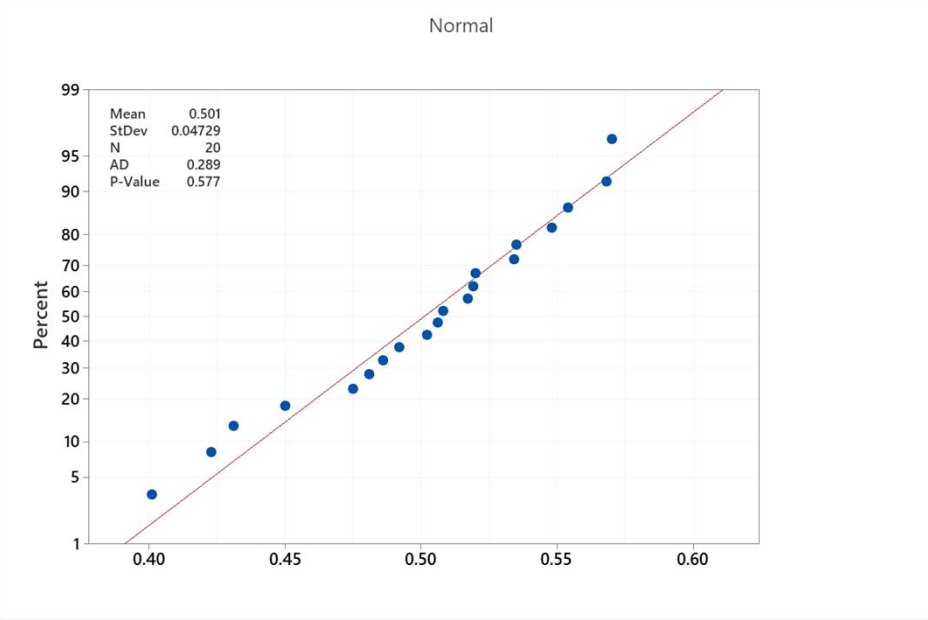
Data set of wood specific gravity data from Draper and Smith (1966) was contaminated by Rousseeuw and Leroy (1987).

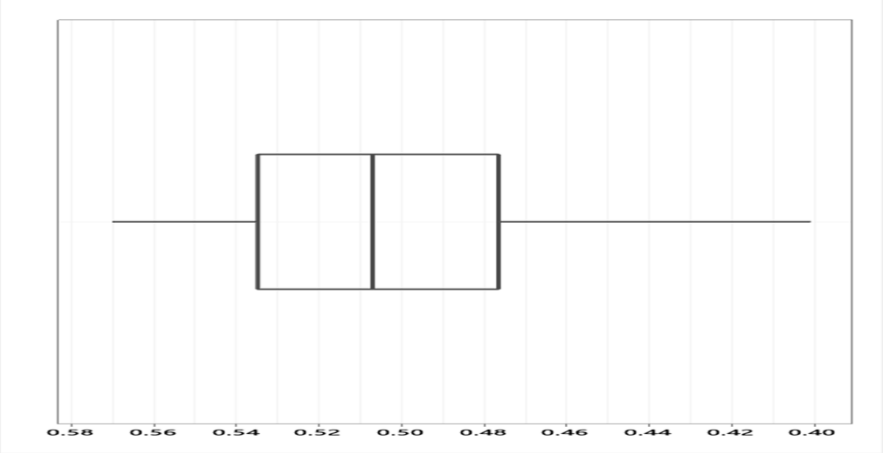
- The data consists of 20 observations.
- The contaminated observations are observations 4, 6, 8 and 19.



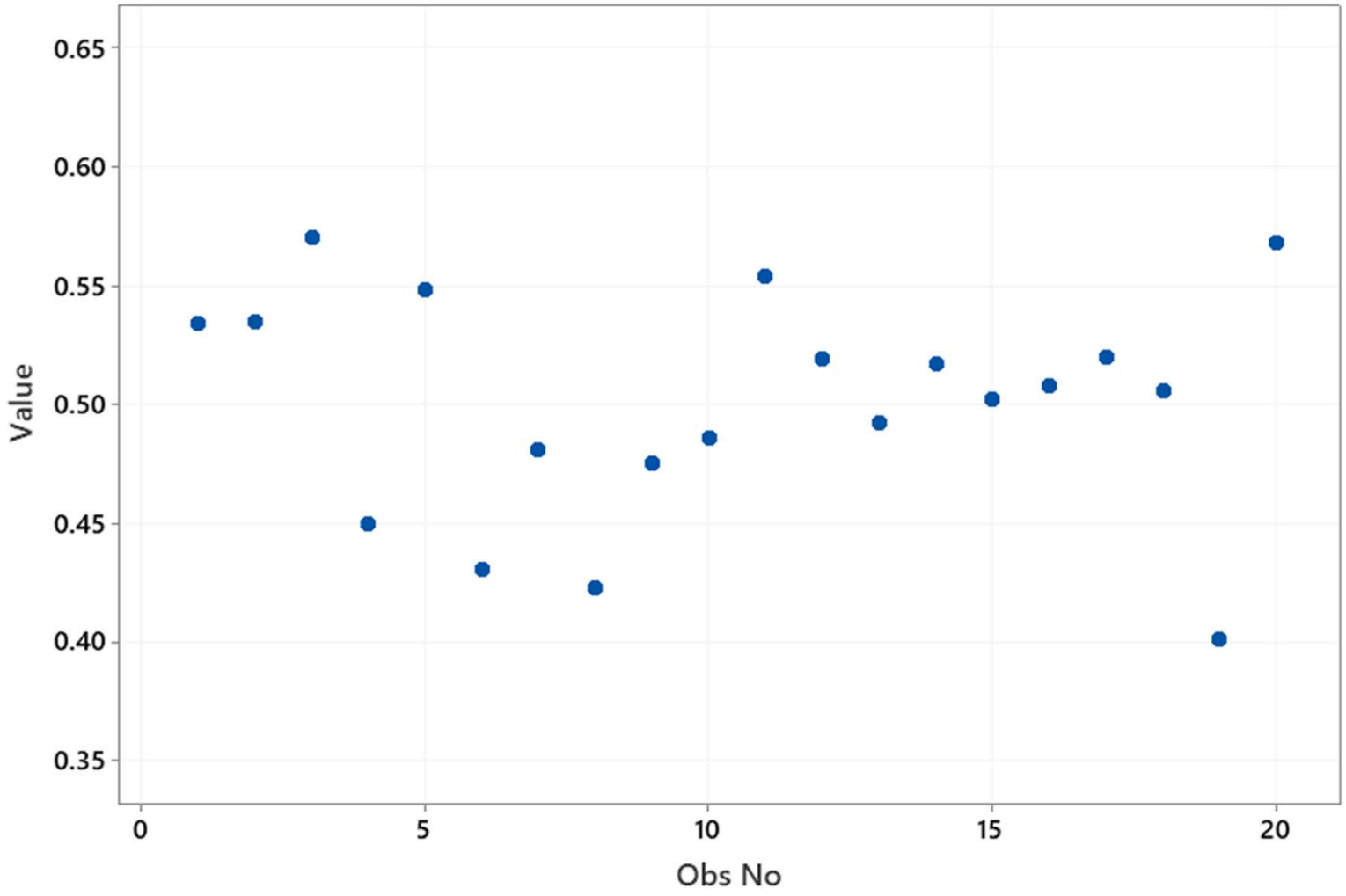
Example Wood specific gravity data

Obs No	1	2	3	**4**	5	**6**	7	**8**	9	10
Value	0.534	0.535	0.57	**0.45**	0.548	**0.431**	0.481	**0.423**	0.475	0.486
Obs No	11	12	13	14	15	16	17	18	**19**	20
Value	0.554	0.519	0.492	0.517	0.502	0.508	0.52	0.506	**0.401**	0.568





Example Wood specific gravity data: How does it work?

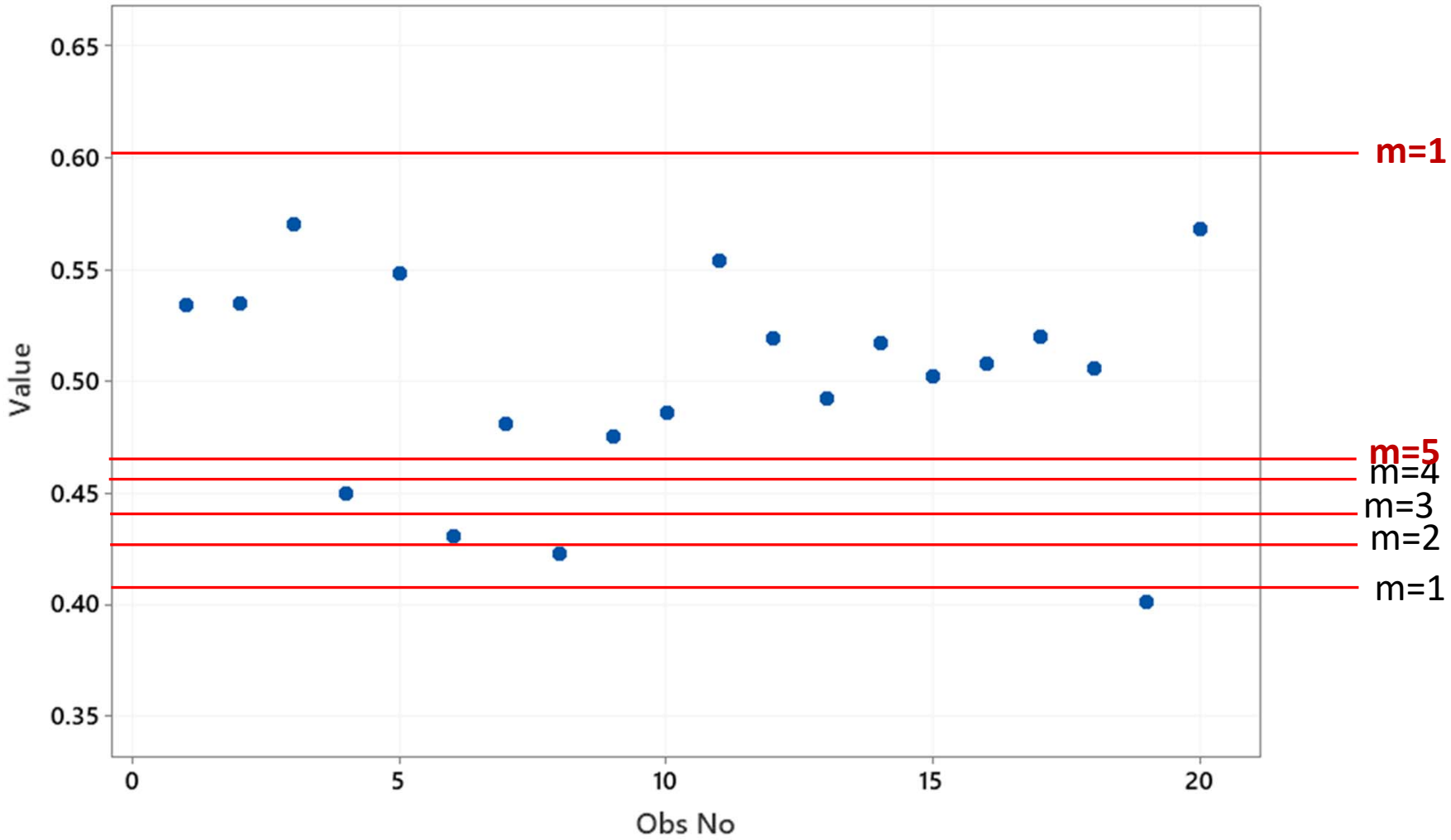


Example Wood specific gravity data: The Fences

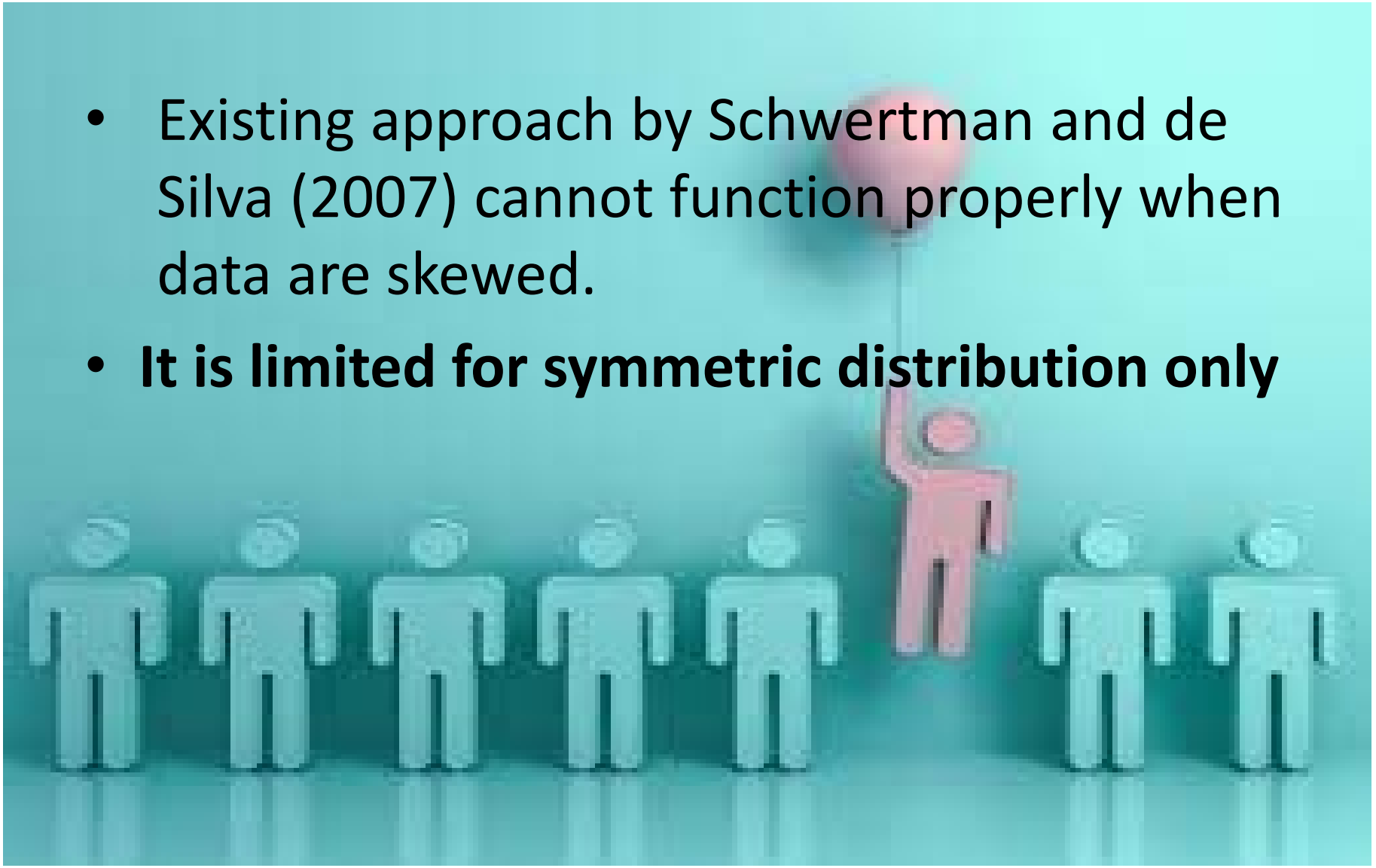
Fences	$C_m = n\,\alpha_{n,m}$	$\alpha_{n,m}$	$t_{df}, \alpha_{n,m}$	Lower	Upper
m=1	0.287682	0.014384	2.38884	**0.4060**	**0.6080**
m=2	0.961279	0.048064	1.76149	**0.4325**	**0.5815**
m=3	1.7273	0.086365	1.42335	**0.4468**	**0.5672**
m=4	2.53532	0.126766	1.18186	**0.4570**	**0.5570**
m=5	3.3686	0.16843	0.98831	**0.4652**	**0.5488**
m=6	4.2192	0.21096	0.82297	**0.4722**	**0.5418**

Example Wood specific gravity data

Fences	m=1	m=2	m=3	m=4	m=5	m=6
Lower	0.406	0.433	0.447	0.457	0.465	0.472
Upper	0.608	0.582	0.567	0.557	0.549	0.542



Problems



- Existing approach by Schwertman and de Silva (2007) cannot function properly when data are skewed.
- **It is limited for symmetric distribution only**

Objectives

Incorporating Skewness into Sequential Fences

- To adjust the coverage of the sequential fences according to the shape of the underlying distribution
- To minimize of misclassifying the regular observations on the long tail as outliers.

Our Contribution

ADJUSTED SEQUENTIAL FENCES FOR DETECTING UNIVARIATE OUTLIERS IN SKEWED DISTRIBUTION

MOMENT BASED MEASURE OF SKEWNESS

- Biased estimator of the population skewness and unbiased skewness is given as

$$MS = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{(N-1)s^3}$$

where s is the standard deviation of the sample and N denoted the sample size.

Using the similar procedures as sequential fences, the proposed method is written as

$$\boxed{ASF = q_2 \pm \frac{t_{df,\alpha_{nm}+MS}}{k_n} IQR}$$

Our Contribution

Wong, H. S. and A. Fitrianto. 2019. Adjusted Sequential Fences for Detecting Univariate Outliers in Skewed Distributions; *ASM Science Journal* , 12 (5), 107-115.

Let’s become GOOD Outlier

Thank You