STATISTICS TEACHER
SUPPORTING THE TEACHING AND LEARNING OF STATISTICS
ASA NCTM

How MAD Must We Be? A Robust Test for Identifying Meaningful Differences Using the Mean Absolute Deviation

Jon Hasenbank and John Appiah Kubi

Grand Valley State University, Allendale, USA

*Corresponding author: hasenbaj@gvsu.edu

1. Introduction

Middle school students' introduction to statistical thinking includes understanding the central role that variability plays in data investigations. In the *Common Core State Standards* (CCSS), which parallel the recommendations of the *Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education II* (GAISE II), sixth graders are introduced to the mean absolute deviation (MAD) as an important tool for measuring variability in quantitative data (*CCSS.6.SP.A.1*, *CCSS.6.SP.A.3*, *CCSS.6.SP.B.4*). The MAD serves as an age-appropriate precursor to the standard deviation; it represents the average distance from the mean and is calculated similarly to the standard deviation:

$$\text{MAD} = \frac{\sum |\text{observed} - \text{mean}|}{n}, \qquad \text{SD} = \sqrt{\frac{\sum (\text{observed} - \text{mean})^2}{n - 1}}$$

We demonstrate in this article how the MAD can be used in ways similar to the standard deviation to determine whether a difference is meaningful.

The seventh grade CCSS standards build on the concepts developed in grade six to engage students in informal statistical inference. Students progress from describing data to understanding how data from samples can be used to estimate population parameters, and they compare samples and discuss whether there is a meaningful difference between them. In doing so, students "assess the degree of visual overlap" between two distributions to make "informal comparative inferences", and they "[express] the difference between the centers as a multiple of a measure of variability" (*CCSS.7.SP.A.2*, *CCSS.7.SP.B.3, and CCSS.7.SP.B.4*). These recommendations are consistent with the GAISE II developmental framework: at Level A, students use graphs to visually compare groups; at Level B, they use informal reasoning to compare groups while noting the uncertainty caused by sample-to-sample

CITATION: Hasenbank, J., & Appiah Kubi, J. (2025). How MAD must we be? A robust test for identifying meaningful differences using the mean absolute deviation. *Statistics Teacher*, Spring 2025 issue. Available online https://www.statisticsteacher.org variability; and at Level C, students learn to use probability to make inferential comparisons and determine the likelihood that observed group differences could be due to chance alone.

In practice, conducting middle school data investigations in which students make "informal comparative inferences" about two data distributions can be challenging because such investigations often lead to cases where the degree of visual overlap and the calculated difference between the centers do not clearly indicate whether the difference is "meaningful" or not. In statistical terms, deciding whether we have a meaningful difference requires evaluating whether the difference is so large that it is unlikely the samples could have been drawn from the same population – in which case, statisticians say they "reject the null hypothesis" in favor of the alternative hypothesis that the source populations must have been different. Unfortunately, precise age-appropriate heuristics for identifying a meaningful difference are largely absent from middle school curricula and statistics education literature. The GAISE II report and the CCSS standards recommend students collect real data through investigations they have helped design. In such investigations, edge cases are bound to arise, which can leave teachers scrambling to provide closure: "Is this difference meaningful? It's hard to say!"

The GAISE II framework outlines a progression by which students learn to quantify such uncertainty. Students use simulations to develop intuition about sampling variation, *p*-values, and confidence intervals before learning to use formal hypothesis tests (e.g., *t*-tests). But because the formal tests are not available to middle school students and their teachers, curriculum authors resort to intuition ("Do you think there is a meaningful difference?") or use canned data sets or contexts that minimize the likelihood of edge cases arising from the analysis. Even though the GAISE II report recommends students learn to embrace ambiguity when learning statistics, and even though anticipating uncertainty is at the very heart of a good statistical question (see *What Is A Statistical Question?* at Census.gov), it can be dissatisfying if too many data investigations conclude with a soft "maybe".

What is a 'Meaningful Difference'?

Before proposing a new heuristic for determining a meaningful difference, let us examine examples from the Common Core State Standards, from the GAISE II report, and from a published *Statistics Teacher* lesson plan, to see how each treats the question of whether a difference is meaningful. [Note: For simplicity, we will use the term "meaningful" as synonymous with "statistically significant", but we should remember that even a small difference can turn out to be statistically significant under the right conditions (e.g., very low within-group variability or very large sample sizes). Whether a difference is truly meaningful (i.e., important) always depends on the underlying context.]

The CCSS middle school standards provide two scenarios suggesting ways students might decide the question of whether a difference between groups is meaningful. Both are nicely grounded in context, but neither provides a clear heuristic; instead, they rely on intuitive phrases like "generally longer" and "noticeable separation":

> [Students will] draw informal comparative inferences about two populations. For example, decide whether the words in a chapter of a seventh-grade science book are generally longer than the words in a chapter of a fourth-grade science book. (*CCSS.7.SP.B.4*)
>
> The mean height of players on the basketball team is 10 cm greater than the mean height of players on the soccer team, about twice the variability (mean absolute deviation) on either team; on a dot plot, the separation between the two distributions of heights is noticeable. (*CCSS.7.SP.B.3*)

Note the hint of a "twice the variability" heuristic in the second example, which we will return to shortly.

The GAISE II report uses language similar to the CCSS, but it provides more detailed examples to illustrate. For instance, two examples of Level B data investigations (Level B is closely aligned with the CCSS middle school standards) are included; both use the context of comparing beak sizes of Medium Ground finches (MGF) and Cactus finches (CF) from the Galapagos Islands. The examples include comparative dot plots and box plots created for two samples (see Figure 1 and Figure 2, respectively):

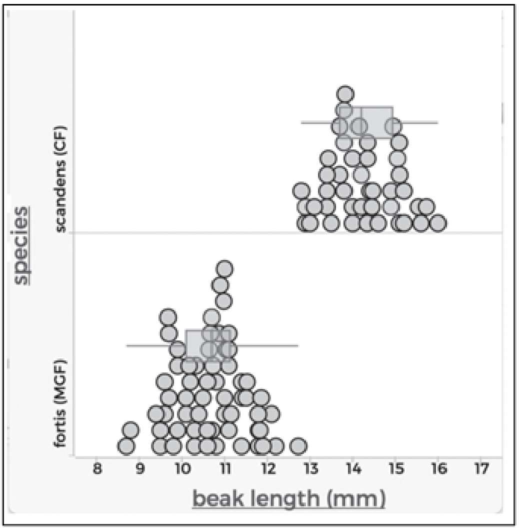

Figure 1. MCF and CF beak length comparison from GAISE II (p. 56)

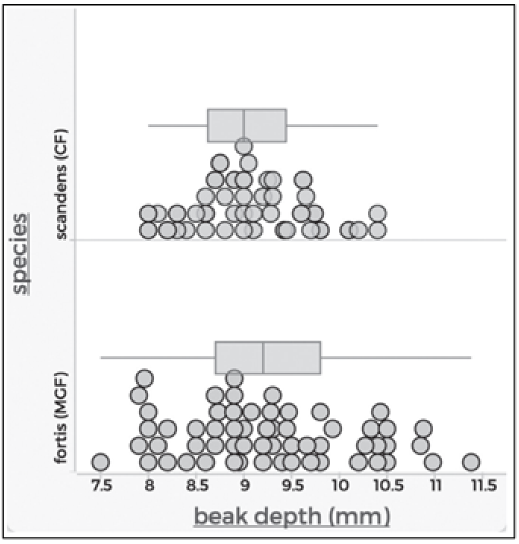

Figure 2. MCF and CF beak depth comparison from GAISE II (p. 58)

As Figure 1 shows, the beak length distributions have almost no overlap between the species (showing a clear difference), whereas the beak depth distributions shown in Figure 2 overlap almost entirely (showing no meaningful difference). In such cases, informal analysis of the group differences is sufficient, as articulated in the sample analyses included in the GAISE II report:

> [For the beak length data:] Considering that the middle 50% of the CF data does not overlap

> with the middle 50% of the MGF data, it would be reasonable to say that in the general finch population, CF beak lengths tend to be greater than MGF beak lengths (p. 56).
>
> [For the beak depth data:] There is no obvious indication as to whether the beak depth of the Medium Ground finches on the Galapagos Islands is greater than or less than the beak depth of the Cactus finches on the Galapagos Islands. The mean beak depth differs by 0.19 mm, which is relatively small for the data given (p. 58).

The CCSS example hints at a "twice the variability" heuristic. The GAISE II examples refer to the middle 50% (the "boxes" of the boxplots) not overlapping. But neither provides concrete guidance that can help teachers and students decide scenarios where there is *partial* overlap.

The "twice the variability" heuristic appears in a more concrete way in the 2023 lesson plan *Exploring Whether a Difference Is a Meaningful Difference* authored by Tim Jacobbe, Christine Franklin, Gary Kader, and Kacie Maddox and published in the *Statistics Teacher*. In it, students collect classroom data to address the question: "Is there a meaningful difference in the number of times students can write their name in 60 seconds with their dominant versus their non-dominant hand?" The teacher's guide includes some sample data and states a clearly articulated "twice the variability" rule for deciding whether the results represent a meaningful difference:

> A common rule is that if the difference between two centers is more than 2 multiples of a measure of variability, then it is a meaningful difference. For example, the difference between two medians should be 2 times the IQR in order to be considered meaningful. The difference between two means should be 2 times the MAD in order to be considered meaningful (p. 7).

The "twice the variability" rule, or *2x rule* for short, can be supported by examining a t distribution with $\alpha = 0.05$ and a large, fixed degrees of freedom. This simplification by the authors has the benefit of avoiding discussions about how sample size affects the sampling distribution of the mean (see this *excellent visualization by Chris Wild*). If we use the standard deviation as our measure of variation, we can see why the *2x rule* is reasonable: the magnitude of $\frac{\bar{x}_1 - \bar{x}_2}{s}$ is less than the two-sample pooled t statistic, $t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{(2/n)}}$, because for $n > 2$, the standard deviation s is greater than the standard error $s\sqrt{(2/n)}$. Moreover, the *2x rule* becomes increasingly conservative as sample size increases: it will identify fewer and fewer meaningful differences than a t-test as the sample size grows.

From this point forward, whenever we use the *2x rule* based on the MAD, we will use the variable k to refer to the ratio $k = \frac{\bar{x}_1 - \bar{x}_2}{MAD}$. The *Meaningful Difference* lesson teacher's guide recommends using the **larger** of the two MADs when expressing the mean difference as a multiple of the MAD. For the sample data provided in the teacher's guide (Figure 3), the means (and MADs) are 23.1 (4.6) and 10.5 (2.2) names per minute for the dominant and non-dominant hands, respectively. We calculate $k = \frac{23.1 - 10.5}{4.6} = 2.7$, and because $k > 2$, we conclude that the difference is meaningful. Indeed, an independent samples t-test confirms that result: at the 95% confidence level, with $df = 24$ and $t = 6.95$, we obtain $p < 0.0001$ (confidence interval = [8.813, 16.26]). But this is an extreme case.

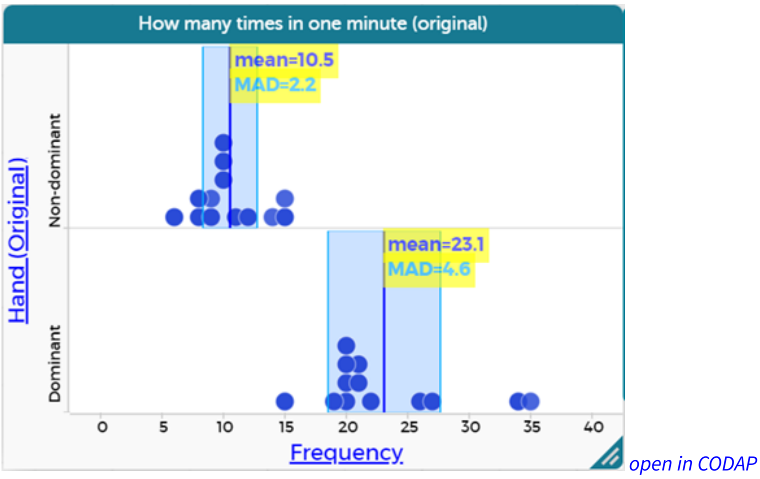

Figure 3. Original data from the *Meaningful Difference* teacher guide

To obtain an edge case, we used CODAP's Scrambler plug-in to shuffle the labels (Dominant vs. Nondominant) for the frequency data in Figure 3, which yielded the data displayed in Figure 4 (*access both data sets in CODAP*). The shuffled data constitutes an edge case because there is evidence of a difference – e.g., the five greatest outcomes were associated with the dominant hand – but there is a fair amount of overlap between the distributions. A t-test returns a statistically significant result ($n = 26, t = 2.44, p = 0.022, 95\% CI = [1.062, 12.63]$), but the difference is not meaningful according to the *2x rule*: $k = \frac{20.2-13.4}{6.7} = 1.01 < 2$. If these results were obtained as part of the *Meaningful Difference* data investigation, we would conclude the difference was not a meaningful one, contrary to the results of the t-test.

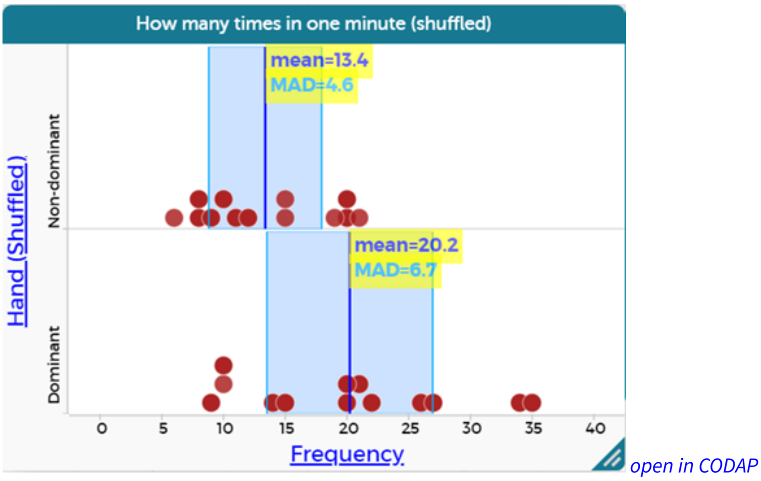

Figure 4. Scrambled data from the *Meaningful Difference* teacher guide

A MAD alternative to the t-test:

To better address the edge cases, we have developed a middle school-appropriate test based on the MAD that improves upon the *2x rule* without requiring knowledge beyond what is expected in the Common Core State Standards for Grades 6 and 7. Teachers and students may use this new "k-test" to reliably decide whether their data investigations have found a meaningful difference.

Where investigations like the *Meaningful Difference* lesson plan run into trouble is when it comes to interpreting the k value, or the difference between the means expressed as a multiple of the MAD (*CCSS.7.SP.B.3*). To handle edge cases, we need to answer the question, "How large must k be for the difference to be statistically significant?" No constant threshold like the *2x rule* will work reliably: a constant threshold would not account for the expected decrease in variability in the sampling distribution of the mean as sample size increases.

It turns out a MAD-based analog to the t distribution exists, though it is not commonly covered in undergraduate statistics texts. It is the H distribution, named for statistician Erna Herrey (pictured in Figure 5), who derived it in her 1965 paper, *Confidence Intervals Based on the Mean Absolute Deviation of a Normal Sample*. Herrey's results, together with a key formula from Revets in *One-norm misfit statistics* (2009), allow the derivation of a formula, $k_{n,\alpha} = t_{2n-2,\alpha} \cdot \sqrt{\pi/(n-1)}$, for calculating the threshold value $k_{n,\alpha}$ corresponding to a statistically significant t-test with sample size n and significance level α. This interactive Google Sheet automates the calculation of the k-critical value for any n and α, and Table 1 shows the particular k-critical values for a handful of common sample sizes at the $\alpha = 0.05$ significance level. For details on our derivation of the k-critical value formula, see the sidebar.

Figure 5. Photo of statistician Erna Herrey (Wikipedia Commons)

Table 1. k-critical values ($\alpha = 0.05$) for select sample sizes

sample size (each sample)	k-critical value
10	1.241
11	1.169
12	1.108
13	1.056
14	1.010
15	0.970
⋮	⋮
20	0.823
30	0.659
60	0.457
120	0.320

For the *Meaningful Difference* sample data, we previously calculated $k = 2.7$, which exceeded the *2x rule* threshold, and we confirmed the difference was significant using a conventional t-test. Table 1 supports the same conclusion: compare $k = 2.7$ with the k-critical value for $n = 13$, $k_{crit} = 1.056$; because $k = 2.7$ exceeds the threshold, the difference is statistically significant. Table 1 reveals how conservative the *2x rule* is, where a gap of just 1.056 MADs is sufficient for $n = 13$ at the $\alpha = 0.05$ statistical significance level.

Recall the calculation of k was based on the conservative choice to use the larger of the two MADs as the denominator. Our derivation of k-critical values does not require making that conservative choice. Instead, we follow Herrey's recommendation to use the average MAD (analogous to a "pooled" standard deviation), rather than the greatest MAD, to calculate k. To see the benefit of using the pooled (average) MAD, consider again the scrambled data from the *Meaningful Difference* lesson (Figure 4). Using the greatest MAD leads to $k = 1.01$, which does not quite exceed $k_{crit} = 1.056$ from Table 1. However, if we calculate k using the pooled MAD, we find $MAD_p = \frac{4.6+6.7}{2} = 5.65$ and $k = \frac{20.2-13.4}{5.65} = 1.20$. This new k-value exceeds the 1.056 threshold from Table 1 and so we can conclude the difference is meaningful, just as the t-test suggested. The use of a pooled MAD allows us to detect smaller significant differences than would be possible if we had to use the maximum MAD.

To test the validity of our k-test algorithm, we used a software package to generate 1,000 pairs of independent random samples from normal populations and recorded how frequently our k-test agreed with an independent samples t-test. The results of three such simulations are shown in Table 2, Table 3, and Table 4, for several combinations of sample sizes and population parameters (note: in the tables, $N(\mu, \sigma)$ refers to a normal distribution with mean μ and standard deviation σ). In each case, the results show agreement rates of 99.5%, 99.7%, and 97.9%, respectively. The few trials in each run where the tests disagree are cases where the p-values closely straddle the significance threshold (e.g., $p = 0.052$ vs. 0.049). We interpret these results as providing further validation of our derived k-test algorithm.

Table 2. t-test vs. k-test simulation, $n = 30$, sampling from $N(0, 2)$: 99.5% agreement with t-test

	***t*-test**	
***k*-test**	**Not significant**	**Significant**
Not significant	956	3
Significant	2	39

Table 3. t-test vs. k-test simulation, $n = 15$, sampling from $N(0, 2)$ and $N(3, 2)$: 99.7% agreement with t-test

	***t*-test**	
***k*-test**	**Not significant**	**Significant**
Not significant	23	1
Significant	2	974

Table 4. t-test vs. k-test simulation, $n = 40$, sampling from $N(0, 2)$ and $N(1, 2)$: 97.9% agreement with t-test

	***t*-test**	
***k*-test**	**Not significant**	**Significant**
Not significant	390	12
Significant	9	589

Sidebar:

In *One-norm misfit statistics* (2009), Revets provides the following key formula relating the standard deviation s to the mean absolute deviation d:

$$s \approx d\sqrt{\frac{\pi}{2}\left(\frac{n}{n-1}\right)}$$

With Revets' formula, we can derive the distribution of

$$k = \frac{\bar{x}_1 - \bar{x}_2}{d_p}$$

from the t-distribution.
Given two independent samples of size n with means $\bar{x}_1$ and $\bar{x}_2$, pooled standard deviation s_p, and pooled MAD

$$d_p = \frac{d_1 + d_2}{2},$$

assuming equal variances, the standard error for the t statistic is given by:

$$SE = s_p\sqrt{\frac{2}{n}}, \quad \text{where} \quad s_p \approx d_p\sqrt{\frac{\pi}{2}\left(\frac{n}{n-1}\right)}$$

Thus,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE} = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p\sqrt{\frac{2}{n}}}$$

$$\approx \frac{(\bar{x}_1 - \bar{x}_2)}{\left(d_p\sqrt{\frac{\pi}{2}\left(\frac{n}{n-1}\right)}\right)\sqrt{\frac{2}{n}}} = \frac{k}{\sqrt{\frac{\pi}{n-1}}}$$

So,

$$t = \frac{k}{\sqrt{\pi/(n-1)}} \quad \Rightarrow \quad k = t \cdot \sqrt{\frac{\pi}{n-1}}$$

Thus, the k-critical values found in Table 1 and in our interactive Google Sheet can be calculated directly from the t-distribution as:

$$k_{n,\alpha} = t_{(2n-2),\alpha} \cdot \sqrt{\frac{\pi}{n-1}}$$

Conclusion

We have presented a robust analog to the independent samples t-test that middle school students and their teachers can use to determine whether their data investigations contain meaningful (statistically significant) differences. The calculation of $k = \frac{\bar{x}_1 - \bar{x}_2}{MAD}$ is precisely what is prescribed in the Common Core State Standards for Grade 7, and our work provides concrete threshold values for determining whether a particular k-value corresponds to a statistically significant difference. We have demonstrated that k-critical values for two samples of size n can be calculated directly from the t-distribution as $k_{n,\alpha} = t_{(2n-2),\alpha} \cdot \sqrt{\frac{\pi}{n-1}}$, allowing us to find precise cut-offs that take sample size into account (in cases where the sample sizes are different, the lesser sample size can be used to identify a conservative threshold).

To be clear, we are not advocating for middle school curricula to begin presenting k-tables like Table 1 to students. Both GAISE II and the CCSS standards focus on developing students' intuition and conceptual understanding about the role of variability in statistical inference. The idea that "samples jump around" – i.e., that small samples tend to vary more from their source populations, as illustrated in *the visualizations by Chris Wild* – is only just emerging at this level. Instead, we propose using the MAD-based $k_{n,\alpha}$ formula to generate k_{crit} values that serve as guidelines for determining meaningful differences when comparing two independent random samples. For example, our formulation lets us say with precision:

- For samples of size 15, our formula gives $k_{\text{crit}} \approx 0.970$. Therefore, a mean difference $(\bar{x}_1 - \bar{x}_2)$ greater than about 1 MAD will make $k = \frac{\bar{x}_1 - \bar{x}_2}{\text{MAD}} > 0.970$, and so should be considered meaningful (i.e., significant at the $\alpha = 0.05$ level).
- For samples of size 30 ($k_{\text{crit}} \approx 0.659$), a mean difference greater than about $\frac{2}{3}$ MAD is meaningful.
- For samples of size 50 ($k_{\text{crit}} \approx 0.502$), a mean difference greater than about $\frac{1}{2}$ MAD is meaningful.

In each case, the MADs may be taken either as the average MAD (typically) or as the greater MAD (conservatively, and especially if one MAD is more than double the other MAD). Thresholds based on other significance levels (e.g., $\alpha = 0.01$) and sample sizes can be easily calculated using the $k_{n,\alpha}$ formula provided above or by using this interactive Google Sheet.

Providing such concrete, sample-size-dependent rules of thumb means students and teachers can bring closure to more of their data investigations, particularly those in which the differences seem "noticeable" but are not extreme enough to pass conservative tests like the *2x rule*. In fact, it is formally correct – though not pedagogically necessary – for middle school teachers to use the phrase "statistically significant" to describe results that contain a "meaningful difference", because our MAD-based test agrees with the corresponding *t*-test in nearly all cases. Moreover, the dependence on sample-size opens the door to early wonderings by students about why using larger samples allows statisticians to classify smaller differences as meaningful. Such wonderings can help bridge the gap between informal and formal statistical inference.

Additional Reading

- Wild, C. (2009, May). *Early Statistical Inferences: "The Eyes Have It"*. USCOTS 2009, Columbus, Ohio. https://www.stat.auckland.ac.nz/~wild/talks/09.USCOTS.html
- Gorard, S. (2015). *Introducing the Mean Absolute Deviation 'Effect' Size. International Journal of Research & Method in Education*, 38(2), 105–114. https://doi.org/10.1080/1743727X.2014.920810