ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2020-74887

An Improved Textbook Rule on the Mean-Median Inequality for Discrete Data

Kesikli Veriler İçin Ortalama-Ortanca Eşitsizliği Üzerine İyileştirilmiş Bir Ders Kitabı Kuralı

Hakan DEMİRTAŞ[a], Ceren VARDAR ACAR[b], Ran GAO[a], İlhan AYDEMİR[c], Ayşegül YABACI[d,e]

[a]University of Illinois Chicago, Division of Epidemiology and Biostatistics, Chicago, USA
[b]Middle East Technical University, Department of Statistics, Ankara, TURKEY
[c]Northern Illinois University, Department of Political Science, Illinois, USA
[d] Bezmialem Vakıf University Faculty of Medicine, Department of Biostatistics and Medical Informatics, İstanbul, TURKEY
[e] Bursa Uludağ University Institute of Health Sciences, Department of Biostatistics, Bursa TURKEY

ABSTRACT Objective: Many introductory statistics books cover the mean-median inequality, which states that if the skewness value is positive/negative, the mean is greater/less than the median. However, this textbook rule is often violated especially when one tail is long and the other is heavy. The purpose of this paper is to propose a refinement that solves the problem to a meaningful extent by bringing the area to the left and right of the median into the picture for discrete data, where violations are more common and severe. The improved version is simple and effective enough to replace the existing rule. **Material and Methods:** Three distributional settings were utilized for illustration: The Poisson, binomial, and discretized normal mixture distributions. A simulation study was devised to assess the relative performances of the current and new rules for count data under the Poisson distribution assumption. **Results:** The new rule adds a simple layer to the current rule: For right skew, the mean is greater/less than the median if the area to the left of the median is less/greater than the area to the right. Similarly, for left skew, the mean is less/greater than the median if the area to the left the median is greater/less than the area to the right. In other words, the new component comes in the form of a comparative area restriction. **Conclusion:** All three distributional examples lead to the same conclusion: The proposed version is associated with substantially better results. Although it is not a complete solution, it is a serious improvement.

Keywords: Tail behavior; introductory statistics; symmetry; mean; median

ÖZET Amaç: Birçok istatistiğe giriş kitabı, çarpıklık değeri pozitif/negatif ise ortalamanın ortancadan büyük/küçük olacağını ifade eden, ortanca-ortalama eşitsizliğini kapsar. Fakat bu kural özellikle bir kuyruk uzun ve diğer kuyruk ağır olduğunda ihlal edilmektedir. Bu çalışmanın amacı, ihlallerin daha yaygın ve şiddetli olduğu kesikli (süreksiz) veriler için, ortancanın sol ve sağ tarafındaki alanı hesaplamaya dahil ederek, problemi anlamlı ölçüde çözen bir iyileştirme önermektir. Geliştirilen kural, mevcut kuralın yerini alacak kadar basit ve etkilidir. **Gereç ve Yöntemler:** Örnek olarak üç dağılım ayarı kullanılmıştır: Poisson, binom ve kesiklileştirilmiş normal karışım dağılımları. Poisson dağılımı varsayımı altında sayma verileri için mevcut ve yeni kuralların göreceli performanslarını değerlendirmek üzere bir simülasyon çalışması tasarlanmıştır. **Bulgular:** Yeni kural mevcut kurala basit bir katman ekler: Sağ çarpıklık için, ortancanın solundaki alan sağındaki alandan daha küçükse/büyükse, ortalama ortancadan daha büyük/küçüktür. Benzer şekilde, sol çarpıklık için, ortancanın solundaki alan sağındaki alandan daha büyükse/küçükse, ortalama ortancadan daha küçük/büyüktür. Başka bir deyişle, yeni bileşen karşılaştırmalı bir alan kısıtlaması şeklinde gelir. **Sonuç:** Her üç dağılımsal örnek de aynı sonuca varmaktadır: Önerilen versiyon büyük ölçüde daha iyi sonuçlar çıkarmıştır. Bu tam bir çözüm olmasa da, ciddi bir gelişmedir.

Anahtar kelimeler: Kuyruk davranışı; istatistiğe giriş; simetri; ortalama; ortanca

Correspondence: Ayşegül YABACI
Bezmialem Vakıf University Faculty of Medicine, Department of Biostatistics and Medical Informatics, İstanbul, TURKEY
E-mail: aysegulyabaci@gmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 13 Mar 2020 ***Received in revised form:*** 01 May 2020 ***Accepted:*** 03 May 2020 ***Available online:*** 28 May 2020

Skewness is a measure of distributional asymmetry of the probability distribution of a real-valued random variable about its mean. It tells us in what way the individual values deviate from the mean. In particular, it indicates if the larger deviations are associated with the larger or smaller values in the data set, corresponding to positive and negative skew, respectively. The qualitative interpretation of skew is based on the reverse direction that the curve appears to be leaning. For a unimodal distribution, positive skew implies that the tail on the right side of the probability density function is longer than the tail on the left side (the mass of the distribution is concentrated on the left), whereas the opposite relationship holds for negative skew. For symmetric distributions, the skewness value is zero. The converse of this statement is not necessarily true, because asymmetries such as one tail being long but thin and the other being short but heavy may even out. Right skewness is more common in real life especially when observed or measured values have an intrinsic lower limit and no obvious upper limit (e.g., cost of living, human weight, time to completion, income), with most data points being closer to the minimum value rather than the maximum. Left skewness is encountered when high values are in majority and low values tail off to the left (e.g., score distribution on an easy exam). In discrete and multimodal continuous distributions, the sign of skewness is difficult to interpret.

A textbook rule often taught in basic statistics courses is that the mean is to the right of the median under positive (right) skew, and to the left of the median under negative (left) skew. It is possible to express this basic premise in different ways: The mean lies toward the direction of skew (the longer tail) relative to the median for skewed distributions;[1] the mean is farther out in the long tail than is the median.[2] The current rule should be regarded as a general trend, not a set-in-stone law since it is prone to violations, with an unacceptably high failure rate.[3] It may not even be fair to call it a rule. This paper is motivated by the need of constructing a new version that represents a significant enhancement.

We go with the most common definition of skew, which is predicated upon the third standardized moment $\nu = E[((X-\mu)/\sigma)^3]$ where X is a random variable with mean μ and standard deviation σ. Some alternatives have also appeared in the literature.[4-6] In theory, it can take any value on the real line, but it is usually smaller than 5 in absolute value in most cases. To estimate skewness from a sample of n observations, one could use a simple estimator $\hat{\nu} = \frac{1}{n}\sum_{i=1}^{n}((x-\bar{x})/s)^3$, $i = 1, \dots, n$ where $\bar{x}$ and s are the sample mean and standard deviation, respectively. An unbiased estimator can be obtained by replacing $\frac{1}{n}$ by $\frac{n}{(n-1)(n-2)}$.[7] Unbiasedness is immaterial for the purposes of this work as we are only interested in the sign of skew as far as the textbook rule of interest is concerned.

Continuous violations are mild and rare, the current rule is good enough for the most part.[4] It is known to hold within the Pearson family. Even under this family, infractions can happen in small samples. Violations are more common and severe for discrete data. That is why we direct our attention to the discrete case, where the median may not (and in majority of cases, does not) split the data in half. Violations occur when the area (the proportion of data points) to the left and right of the median are unequal in a certain way. For right skew, we typically detect a violation when the area to the left of the median is larger than the area to the right, because one tail is long (the right tail) and the other tail (the left one) is heavy. The opposite connection is valid for left skewness. The remainder of the paper focuses on formulating a new criterion for the discrete case, taking this observation into consideration. The crux of the proposed rule is the relative magnitudes of the proportion of data to the left and right of the median. We capitalize on the direction of this area inequality for a refined variation of the current textbook rule.

MATERIAL AND METHODS

Let us re-iterate the current rule: For right/left-skewed data, the mean is greater/less than the median. This makes a great deal of sense to a beginner since its conceptualization, visualization, and contextualization are straightforward. The problem is that it can easily get violated for discrete data.

The new rule adds a simple layer to the current rule: For right skew, the mean is greater/less than the median if the area to the left of the median is less/greater than the area to the right. Similarly, for left skew, the mean is less/greater than the median if the area to the left the median is greater/less than the area to the right. In other words, the new component comes in the form of a comparative area restriction. When this restriction is made a part of the rule in an attempt to improve the process, by and large the direction of the mean-median inequality corrects itself. To re-express this symbolically, let μ, m and Δ be the mean, median, and difference between the area to the left (A_L) and right (A_R). For $\nu > 0,\ \mu > m$ if $\Delta = A_L - A_R < 0$ and $\mu < m$ if $\Delta > 0$. Similarly, for $\nu < 0,\ \mu < m$ if $\Delta > 0$, and $\mu > m$ if $\Delta < 0$. This expansion fixes a solid portion of violations. The downside is that it creates a new violation albeit to a much lesser extent. The area argument does not offer a perfect solution, but it substantially reduces the number of violations in a convincing manner, as demonstrated below.

Three distributional settings were utilized for illustration: The Poisson, binomial, and discretized normal mixture distributions. The Poisson distribution with the rate parameter $\lambda > 0$ is a great place to start testing the suggested rule. Although skew is always positive ($\nu = 1/\sqrt{\lambda}$), the mean becomes less than the median if the decimal part of the rate parameter is greater than $log2 \approx 0.693$.[3] More specifically, the median is approximately $\left\lfloor \lambda + {}^{1}/_{3} - {}^{0.02}/_{\lambda} \right\rfloor$ where $\lfloor y \rfloor$ is the largest integer that is less than y.[8] A simulation study was devised to assess the relative performances of the current and new rules for count data under the Poisson distribution assumption. The integer part of the rate parameter was chosen to be 0, 2 and 5 and the decimal part was assumed to vary between 0.50 and 0.89 with increments of 0.01 (40 different values for each of the three ranges that only differ in the integer portion are denoted by R_1, R_2 and R_3; these lower and upper bounds in the decimal range were selected so that they almost equidistant from 0.693) Three sets of sample sizes $n = (n_1, n_2, n_3)$= (200,1000,10000) were used, leading to 3*3=9 sample size-range combinations. The experiment was performed for N=100 simulation replicates for each λ value, making the total number of simulated data sets 100 x 40 x 9=36000 whose sample means and medians were calculated and compared, and whether or not violations occurred with respect to the current and new rules were determined.

RESULTS

Figure 1 is concerned with 4000 points that come out of the $n_1 R_1$ combination. The points on the left and right parts correspond to legitimate cases and violations, respectively, by to the current rule, with the area difference Δ on the y axis. Violations go away with the introduction of area condition (1948 violations are removed, corresponding to 48.7% of cases). The disadvantage is that a new sort of violation ensues, but it carries far less negative connotations: Some points with a positive Δ (positive differences on the left part of the graph) go from legitimate to violated. The magnitude of the problem is much smaller though (26 new violations are emerged, corresponding to 0.65% of cases). This complication becomes nonexistent in larger samples, as we will see shortly.

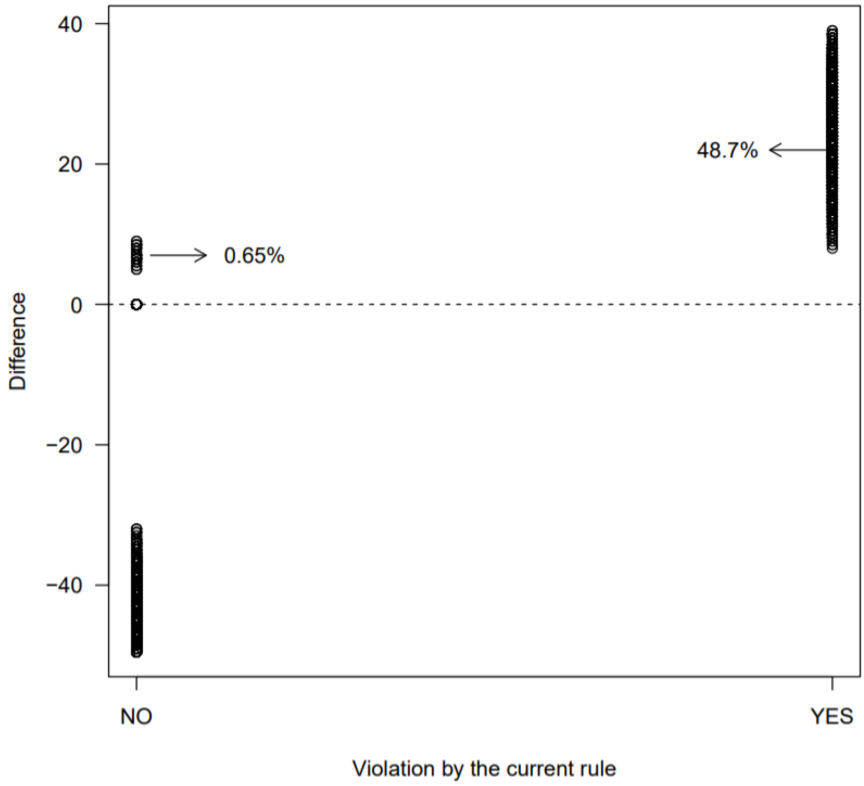

FIGURE 1: The area difference versus the violation status by the current rule.

Given the sign of skew, connecting the area difference to the mean-median inequality removes an overwhelming majority of the violations with a minor glitch. The good news is that the benefits of the new rule far outweigh the losses, giving rise to an effectively working system. Figure 2 shows how well the improved version operates, along with a clearer demonstration of the above-mentioned problem across six range-sample size combinations, in which $n = n_1$ or n_2. The area differences are sorted across the violation status by the current rule. Taken as a matrix plot, rows and columns stand for the sample size and range, respectively. The thick curve in plot [1, 1] (n_1R_1) represent the violated cases under the current rule, where the area difference is positive for a right-skewed distribution. The new rule moves these points to the valid territory. The points on the regular curve denote legitimate cases under the current rule. They mostly fall into the region where the area difference is negative; the ones that are above the x axis are the new violations. In plots [1, 2] (n_1R_2) and [1, 3] (n_1R_3), we see the same trend (a large number of violations under the current rule are removed in exchange for a small number of new violations under the new rule) to a slightly worse degree. Another minor setback that can be seen in plot [1,3] is that a very small number (only three to be exact) of current rule violations remain (the ones with a negative area difference). We see a spotless trend in the second row of the plots that correspond to $n = n_2$: All violations are contained, and no new violations are introduced. The large sample behavior $(n = n_3)$ is no different from this: Violations go away completely. The violation rate across all nine sample size-range scenarios are graphed in Figure 3, where the thin curve stands for the current rule. The violation percentage is fluctuating around 50% (ranging from 45% to 55%, which is an annoyingly high proportion) under the current rule, whereas it goes down significantly for $n = n_1$ (0.65%, 1.97% and 5.52% for R_1, R_2 and R_3, respectively; naturally in harmony with Figure 2 trends) and entirely vanishes when n is larger, with the new adjustment.

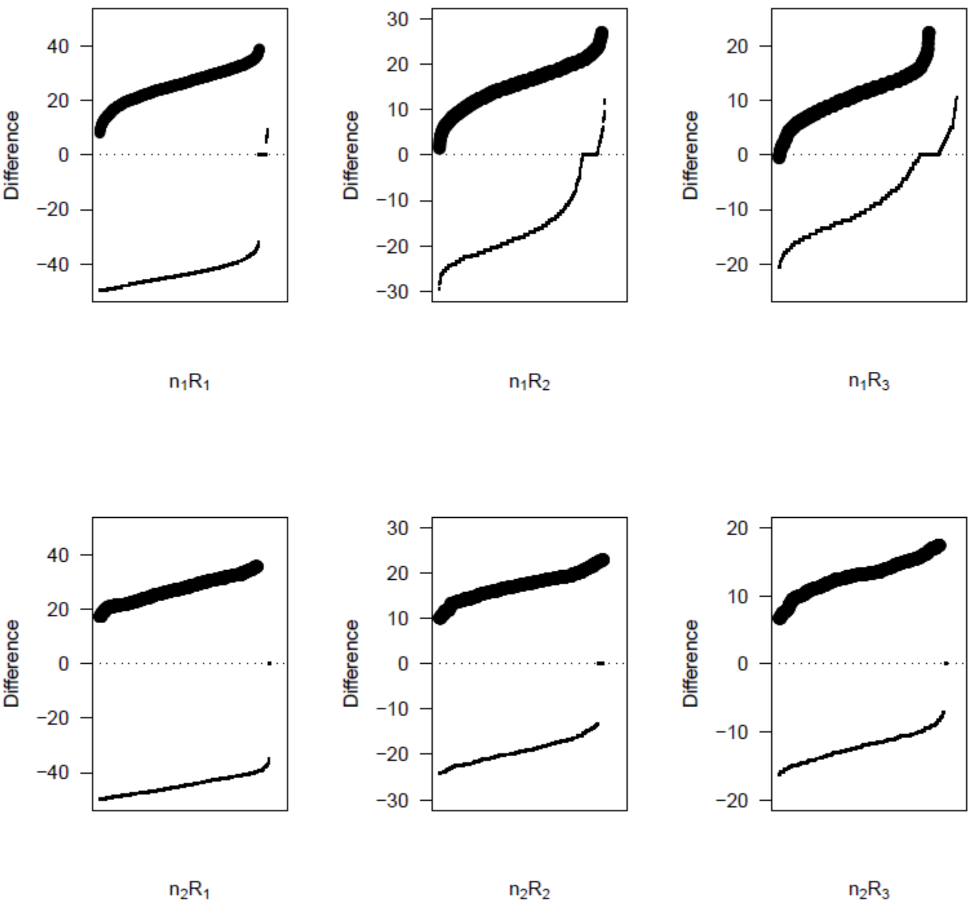

FIGURE 2: The sorted area difference with respect to the violation status via the current rule across six sample size-range combinations, where the thick curve represents removed violations by the new rule.

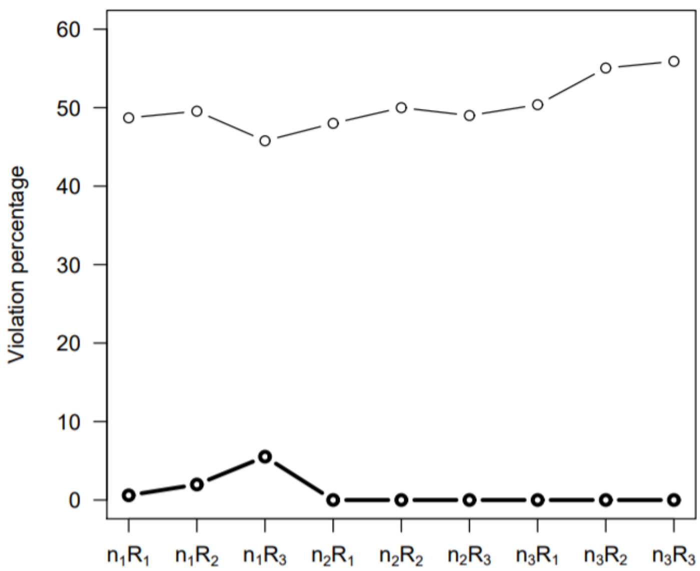

FIGURE 3: An overall comparison graph in the Poisson example regarding the per cent violation for nine sample size-range combinations.

Our second illustration is drawn from the binomial distribution with the sample size (n) and proportion (p) parameters, which could be left or right-skewed depending on the value of p (notice the slight change in notation, in correspondence with the established convention, here n is the sample size parameter, the number of data points is now $nrep$). The skewness expression is given $\nu = \frac{1-2p}{\sqrt{np(1-p)}}$. Skew is negative for $p > 0.5$ and positive for $p < 0.5$. If np is an integer, then the mean (μ) and median (m) coincide. The median is equal to $round(np)$ in cases when either $p \leq 1 - log2$ or $p \geq log2$ or $|m - np| \leq \min(p, 1-p)$ except for the case when $p = 0.5$ and n is odd.[9] Three sample sizes n= (2, 3, 4) and 81 values for p, ranging from 0.10 to 0.90 with increments of 0.01 with $nrep = 50$ and $N = 100$ simulation replicates for each (n, p) combination were harnessed. Figure 4 is a 2 x 2 graph that shows the per cent violation across 8100 simulated data sets for each of the three specific n's as well as the aggregated one, under the current (regular curve) and new (thick curve) rules. In the upper left graph (*n=2*), the standard (current) rule fails miserably especially when $0.4 \leq p \leq 0.6$. Here is why: When $p = 0.4$, $\nu > 0$, $\mu = np = 0.8$ and $m = 1$ because $|m - 0.8| \leq 0.4$ by Kaas and Buhrman (1980). It is a violation because $\mu < m$ when $\nu > 0$. Similarly, when $p = 0.6$, $\mu > m$ although $\nu < 0$, $\mu = np = 1.2$, and $m = 1$ because $|m - 1.2| \leq 0.4$. When $p = 0.5$, $\mu = m$ the sample mean can easily move from one side to another. In the upper right graph ($n = 3$), we see a severe violation by the current rule around $p = 0.27$ and $p = 0.73$ by the same logic. The lower left graph ($n = 4$) tells a fairly compatible story: the mean is on the wrong side of the median for the values around $p = 0.21$, 0.45, 0.55 and 0.79. Incorporating the area restriction to the rule yields much fewer violations for individual levels of all three *n*'s (and not surprisingly the overall picture, see the lower right graph). The new variation's performance is markedly better, suggesting that the new rule works properly regardless of the sign of skew.

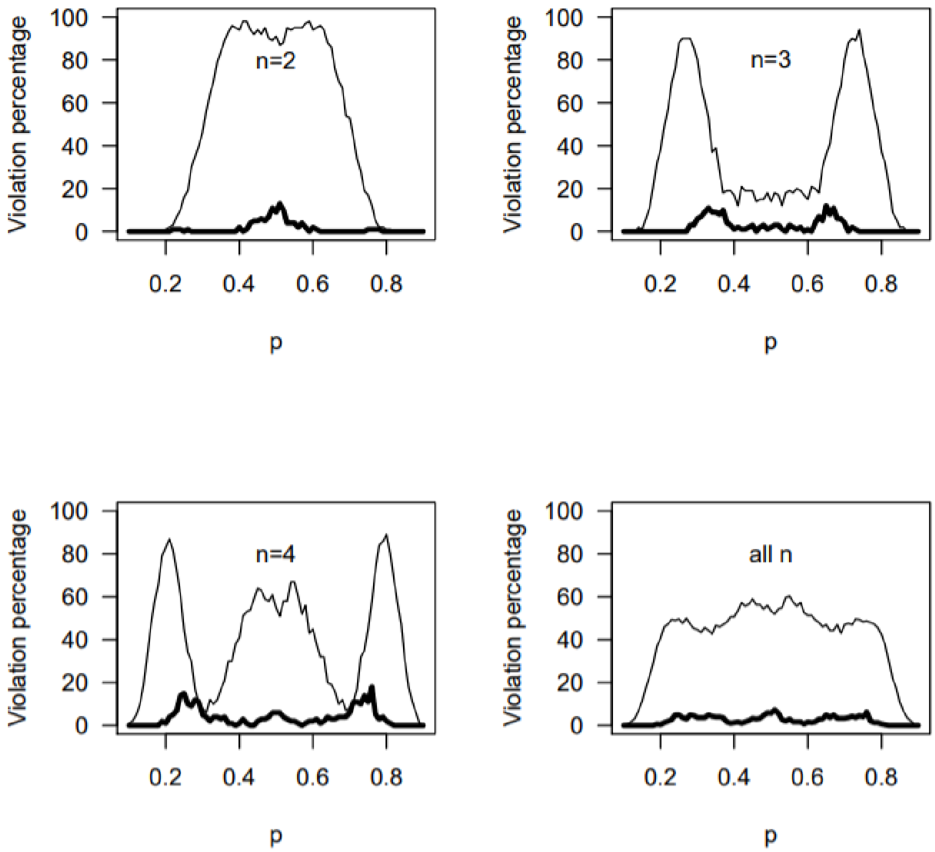

FIGURE 4: The violation percentages by the current and new rules for three different n values, and the aggregated version in the binomial example.

For multimodal distributions, the direction of skew is not immediately obvious by visual inspection of the underlying shape. To gauge the plausibility of the improved rule outside the realm of unimodal distributions, we employ a normal mixture density before rounding the variates to the nearest integer for four bimodal situations. A normal mixture distribution can be formulated as $\pi N(\mu_1, \sigma_1^2) + (1 - \pi)N(\mu_2, \sigma_2^2)$ with the mixing proportion π. Depending on the choice of parameter values, it could take different shapes. As can be seen in Figure 5, $(\pi, \mu_1, \sigma_1, \mu_2, \sigma_2)$ are chosen to be (0.5,0,1,3,1) (upper left), (0.6,0,0.5,2,0.5) (upper right), (0.3,0,1,3,1) (lower left), and (0.5,0,1,4,1) (lower right) that correspond to different bimodal settings with respect to the height of the modes and the distances between them. The characteristics of the discretized data are obviously different, but the bimodal shapes are not distorted. The number of data points (n) are set to range from 20 to 2000 with increments of 20 (100 different values with N=100 simulation replicates for each level of n). In Figure 6, we plot the violation percentages by the current (regular curves) and new rules (thick curves) for each set of specifications. The new rule is the clear winner in all four cases. In addition, most statistical procedures give more stable and predictable results as n gets larger, and the suggested rule reflects this general tendency, whereas the current rule shows erratic, idiosyncratic behavior. With a few exceptions (when n is small), the suggested variation turns out to be uniformly superior to the existing rule. The number of violations becomes either nonexistent (plot [1,2]) or negligibly small (plots [1,1] and [2,1]) or acceptably low (plot [2,2]).

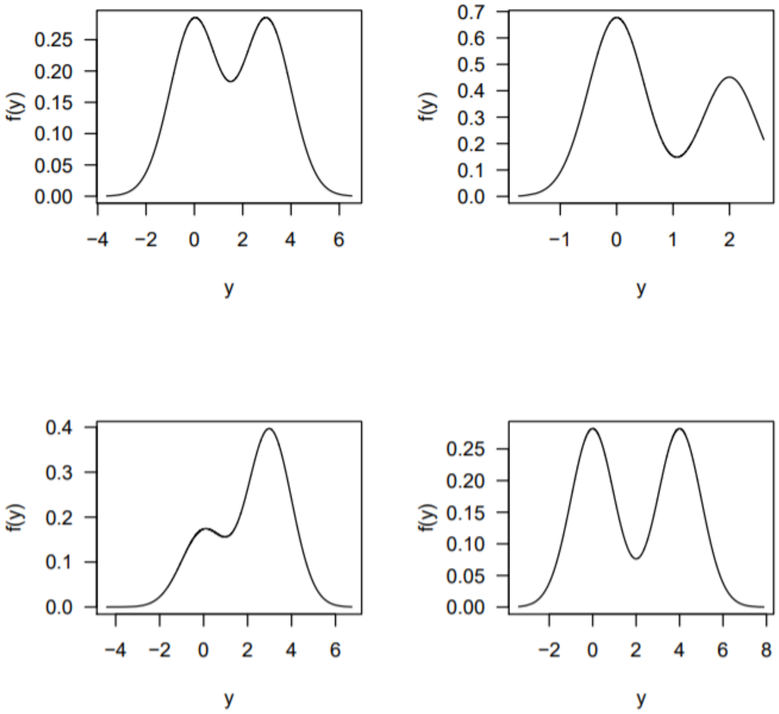

FIGURE 5: Four bimodal shapes of the normal mixture distribution.

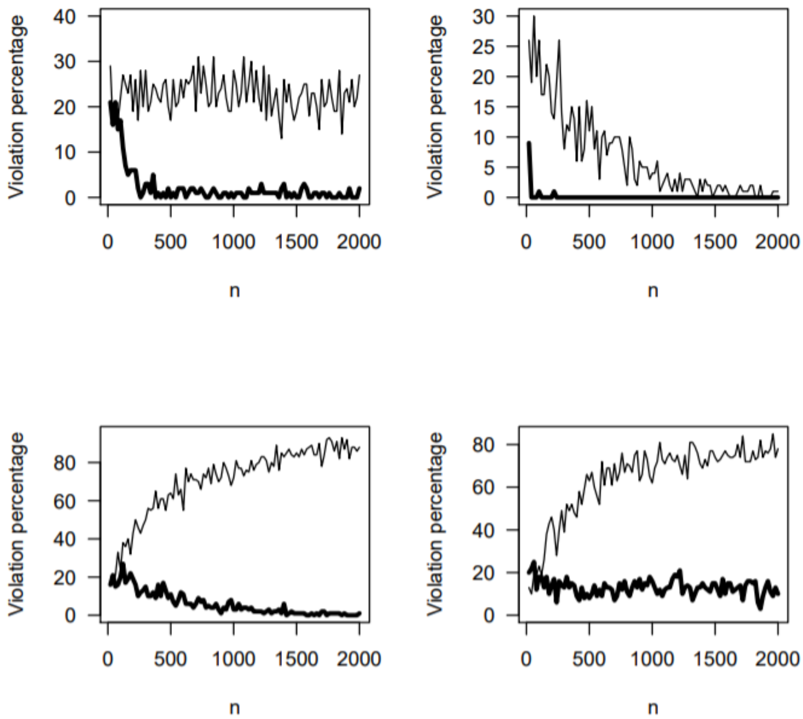

FIGURE 6: Comparison of the violation percentages for 100 different sample sizes in the discretized normal mixture example.

All three distributional examples lead to the same conclusion: The proposed version is associated with substantially better results. No, it is not a complete solution, but it is a serious improvement.

The reported results are reproducible and the R code used in this manuscript is publicly accessible at http://demirtas.people.uic.edu/computer_code_mean_median_paper.txt [10]

DISCUSSION

Here are some points that deserve mention in no particular order: (a) While the mode is occasionally presented as a part of the current rule in some books[11] -the median is being between the mean and the mode-, it is out of the scope of this paper since the mode is not the gist of the textbook rule despite its importance as a statistical quantity. This manuscript concentrates on the mean-median inequality, and no attention is paid to the mode. (b) The area to the left or right may be zero if the median is sitting on an extreme category, but it does not change the nature or effectiveness of the new rule, which hinges upon the difference. (c) Skew is undefined for some distributions (e.g., Cauchy), and it may be infinite (e.g., $f(x) = x^{-2}I[x > 1]$). Furthermore, it could be very sensitive to outlying data points as it involves cubes, but these complications equally apply to the current and proposed rules. (d) The sign of the sample skew may not conform to what the theory suggests (e.g., a sample that is based on Poisson distribution may have a negative skew especially when the number of data points is small), but this does not invalidate or affect the new rule in any way. (e) One can argue that the mean may not always be a sensible measure for discrete data (e.g., ordinal data). This is true, but again, it is not a problem that is originated by the new rule. (f) The median usually does not split the data in half for discrete data (ironically, this is the engine of the new rule), so the way we interpret it is a bit different from the continuous case. Nevertheless, no matter how we look at it, the median is still the number $in\ the\ middle$ with or without ties. (g) Monotonic skew-reducing transformations (e.g., square-root, logarithmic) do not change the relative positions of the mean and median, so they are irrelevant to this work. (h) The use of alternative skew formulas does not provide any extra utility in terms of reducing the discrepancy between what the current rule suggests and reality. (i) Generalizability of the presented results may be considered doubtful by some in the absence of a mathematically cogent proof. However, the outcomes are suggestive and compelling enough for us to take a point of advocacy for the enhanced rule. The proposed improvement is not tied to any distributional shape or sample size assumptions, which constitutes a major plus in terms of broad applicability across a rich spectrum of discrete data contexts.

As discussed before, violations in the continuous case are mild; distortions on the comparative magnitudes of the mean and median are typically indiscernible. Nevertheless, the problem could be solved by incorporating left and right kurtosis (computed for the two halves of the data), reflecting the peakedness or elongation behavior at the tails.[12] Given the low severity and frequency of the problem, trying to come up with a corrective mechanism may not be worth the effort, and it would probably be too complicated to cover at a beginner's level course. Of note, the relative superiority of the new rule becomes more pronounced if the number of categories is small. If it is large, a discrete distribution nearly becomes an approximation to a continuous one, producing not so many violations under the existing rule.

The proposed rule is not meant to be miraculous or free of violations, as it does not always work; counter examples can be found (as mentioned in Section 2) or constructed. It is difficult to imagine that there exists a rule that works well in all situations including the small sample cases in this context. However, it is a considerable improvement over the current rule, as shown in a few distributional settings. The newly added component is the difference between the areas to the left and right of the median. It is conceptually and computationally straightforward enough to be the new textbook rule. Formulating a better guideline concerning the mean-median inequality is useful to address a frequently encountered violation in practice; it has some significance in teaching of elementary statistics courses.

CONCLUSION

We can go bold and embrace a larger augmentation strategy by combining these two rules, which could be stated as (a) Go with the current rule, and (b) If *and when* it fails, go with the new rule. This pairing gives nearly violation-proof results; it solves the problem without creating a new one. The two-stage rule may feel a little contrived and unnatural, however, it is hard to contradict the assertion that says *whatever works is optimal*. On a more positive note, we believe that the proposed refinement has potential to be adopted by the authors of the introductory books, with or without this sequential angle.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

All authors contributed equally while this study preparing.

REFERENCES

1. Agresti A, Finlay B. Statistical Methods for the Social Sciences. 4th ed. Upper Saddle River, NJ: Prentice Hall; 2008. p.624.
2. Moore D, McCabe G, Craig B. Introduction to the Practice of Statistics. 8th ed. San Francisco: W. H. Freeman; 2016. p.694.
3. von Hippel PT. Mean, median, and skew: correcting a textbook rule. J Stat Educ. 2005;13(2). [Crossref]
4. Groeneveld RA, Meeden G. Measuring skewness and kurtosis. Journal of the Royal Statistical Society: Series D (The Statistician). 1984;33(4):391-9. [Crossref]
5. Hinkley DV. On power transformations to symmetry. Biometrika. 1975;62(1):101-11. [Crossref]
6. MacGillivray H. Skewness and asymmetry: measures and orderings. The Annals of Statistics. 1986;14(3):994-1011. [Crossref]
7. Rose C, Smith MD. Mathematical Statistics with Mathematica. 1st ed. New York: Springer; 2002. p.481. [Crossref]
8. Choi KP. On the medians of gamma distributions and an equation of Ramanujan. Proceedings of the American Mathematical Society. 1994;121(1):245-51. [Crossref]
9. Kaas R, Buhrman JM. Mean, median and mode in binomial distributions. Statistica Neerlandica. 1980;34(1):13-8. [Crossref]
10. R Development Core Team. R: A Language and Environment for Statistical Computing. Foundation for Statistical Computing Vienna, Austria; 2020. [Link]
11. Levin J, Fox JA, Forde DR. Elementary Statistics in Social Research. 3rd ed. Boston: Pearson Education Group; 2013. p.360.
12. Zenga, M. La curtosi. A new approach to kurtosis. Statistica. 1996;56:87-102.