On Some Alternative Probability Density Metrics for Analyzing Empirical Datasets

Sidney Klawansky[1], Brielle Balswick[2], Meaad Alsayel[2], Iryna Charvachidze[2], Anuka Manghwani[1,2], Pearl Almeida[2], Dharmit Dalvi[2], Janvi Vora[2], and Eugene Pinsky[2(✉)]

[1] Department of Health Policy and Management, Harvard Chan School of Public Health 2, 677 Huntington Ave, Boston, MA 02115, USA
sidney.klawansky@alum.mit.edu

[2] Department of Computer Science, Metropolitan College, Boston University, 1010 Commonwealth Avenue, Boston, MA 02215, USA
{brielleb,sayelmm,irynach,palmeida,dharmit,epinsky}@bu.edu

Abstract. We propose a simple set of probability density shape metrics with intuitive interpretability and complement the Classical statistical metrics of Variance, Skewness, and Kurtosis. These Classical metrics involve squaring of deviations and computation of third and fourth moments. Therefore, they may be overly sensitive to outliers. Therefore, we take The Mean Deviation around the mean, rather than the standard deviation, as the primary measure of data dispersion. This work presents some of our initial results using Mean Deviation and the new metrics of Tailness (an analog of Kurtosis) and Asymmetry (an analog of Skewness). These new metrics use only first and second moments. They have simple interpretations and directly allow us to compare datasets with different measurement units. As such, they give us additional tools for data analysis. We illustrate the proposed metrics for several public datasets.

Keywords: mean absolute deviation · skewness · asymmetry · kurtosis · tailness

1 Introduction

Standard deviation, skewness, and kurtosis are widely used measures in statistical analysis. In computing σ, we use the squares of the distances from the mean μ. For skewness and kurtosis, we use higher-order moments. As noted by Pham in [1], using the L_2 norm is convenient in differentiation, estimation, and optimization. The additive property of variance σ^2 for independent variables is also cited as one of the prime reasons for using the L_2 norm.

In recent years, there has been an increased interest in using the L_1 metrics in data analysis (e.g. [2–6]). Using the L_1 norm is not new. The L_1 norm was considered independently by both Boscovitch and Laplace as early as the 18th

T. Zlateva and G. Tuparov (Eds.): CSECS 2023, LNICST 514, pp. 40–47, 2023.
https://doi.org/10.1007/978-3-031-44668-9_3

century. A historical survey using the L_1 norm is presented in [7,8] and a survey of more recent results is given in [1].

Our starting point is the mean absolute deviation defined as follows: Consider a real-valued random variable X on a sample space $\Omega \subseteq R$ with density $f(x)$, finite mean μ, and cumulative distribution function $F(x)$. If X is a discrete random variable, then Ω is some countable sample space, and $f(x)$ is the probability mass function (or discrete density function).

For any a, we define the mean absolute deviation of X from a as

$$d(X, a) = \mathrm{E}(|X - a|) = \int_{\Omega} \big|x - a\big| f(x)\, dx \tag{1}$$

If $a = \mu$, then $d(X, \mu)$ is the mean absolute deviation from the mean μ. If we take $a = M$, then $d(X, M)$ is the mean absolute deviation from the median. Both of these are denoted as MAD (mean absolute deviation) in the statistical literature, leading to some confusion [1]. In this paper, we use MAD to denote **mean absolute deviation** from the mean and denote it by d. A detailed discussion of mean absolute deviation from the median is presented in [9].

The measure d can be interpreted as the average distance of values of X to the mean μ. It is well-known that $d \leq \sigma$ for any distribution. Therefore, one can expect metrics based on d to be less sensitive to outliers and therefore more robust.

Note that we can remove the absolute magnitude sign in (1) and derive d by computing contributions from points less than or equal to μ and points greater than μ:

$$d = d_- + d_+, \qquad d_- = \int_{x \leq \mu} (\mu - x) f(x)\, dx, \qquad d_+ = \int_{x > \mu} (x - \mu) f(x)\, dx \tag{2}$$

By separately computing the integrals over the left and right sub-spaces, we have eliminated the absolute sign. Similarly, although σ^2 is usually thought of as a single number, for the proposed metrics we explicitly compute contributions to variance from the left ($X \leq \mu$) and right ($X > \mu$) sub-spaces.

$$\sigma^2 = \sigma_-^2 + \sigma_+^2, \qquad \sigma_-^2 = \int_{x \leq \mu} (\mu - x)^2 f(x)\, dx, \qquad \sigma_+^2 = \int_{x > \mu} (x - \mu)^2 f(x)\, dx \tag{3}$$

These contributions from the left and right subspaces in (2) and (3) must be computed separately. As a result, they form the building blocks of the more transparent asymmetry and tailness metrics proposed below. In this paper, we define the following Alternative Metrics:

1. Mean deviation d - instead of standard deviation σ
2. Tailness $T = 2\sigma^2/d^2$ - instead of kurtosis $K = \mathrm{E}((X - \mu)/\sigma)^4$
3. Asymmetry $A = (\sigma_+^2 - \sigma_-^2)/\sigma^2$ - instead of skewness $S = \mathrm{E}(((X - \mu)/\sigma)^3)$

Let us make a few quick observations. The mean absolute deviation d always satisfies $d \leq \sigma$. For example, in Gaussian distributions $d = \sigma\sqrt{2/\pi}$ or about

20% lower than σ. For tailness T, we normalize the variance σ^2 by $d^2/2$ to bring the value $T = \pi$ for the Gaussian distribution close to the kurtosis value $K = 3$. Finally, for asymmetry A, we can show that, unlike Skewness that can be unbounded, Asymmetry is always in the range $-1 \leq A \leq 1$. For this reason, the proposed measure of Asymmetry allows us to compare datasets across different measurement units.

The above definitions for Mean Deviation d, Asymmetry, A, and Tailness, T, allow the computation of these measures without calculating 3^{rd} or 4^{th} moments. Many fields including medicine, epidemiology, psychology, economics, finance, and biology are heavy users of statistical tools. We are optimistic that the proposed metrics can provide a self-consistent system of measurement that will make more results in these fields easier to interpret.

2 Examples: Gaussian, Laplace and Exponential

In this section, we present these Alternative Metrics the for some well known distributions. We start with the normal distribution. The standard representation of the Normal or Gaussian distribution is given in Eq. (4)

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{4}$$

For this distribution, the relationship of d to standard deviation σ, is given by [10]

$$d = \sqrt{\frac{2}{\pi}}\,\sigma \qquad \text{or} \qquad \sigma = \sqrt{\frac{\pi}{2}}\,d \tag{5}$$

For the Normal Distribution, $d \approx 0.798\sigma$. Using (5), we re-express the Normal Distribution as a function of d rather than σ.

$$f(x, \mu, d) = \frac{1}{\pi d} \exp\left(-\frac{(x-\mu)^2}{\pi d^2}\right) \tag{6}$$

The representation of the Gaussian distribution using the d in (6) is simpler than the representation using the σ in (4) in that there is no square root in the factor multiplying the exponential. We also note that in Equation (5) for the Gaussian, when we square both sides, we obtain $\sigma^2/d^2 = \pi/2$. This relationship led to the definition of Tailness as discussed previously.

Similarly, for Laplace and exponential distribution, we have

1. Laplace: mean deviation $d = b$ [11]

$$\text{Classical:} \qquad f(x, \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

$$\text{Alternative:} \qquad f(x, \mu, d) = \frac{1}{2d} \exp\left(-\frac{|x-\mu|}{d}\right)$$

This distribution is symmetric and therefore $\sigma_-^2 = \sigma_+^2$. The asymmetry is therefore $A = 0$ and its tailness $T = 4$. The classical kurtosis for this distribution is $K = 6$.

2. Exponential: density is $f(x) = \lambda e^{-\lambda x}$ with $x > 0$. The standard deviation is $\sigma = 1/\lambda$ and the mean deviation $d = 2/(e\lambda)$ [11]

$$\text{Classical:} \qquad f(x,\sigma) = \frac{1}{\sigma}\exp\left(-\frac{x}{\sigma}\right)$$

$$\text{Alternative:} \qquad f(x,d) = \frac{2}{ed}\exp\left(-\frac{2x}{ed}\right)$$

This distribution is not symmetric and, therefore, $\sigma_-^2 \neq \sigma_+^2$. We compute σ_-^2 and σ_+^2 as follows. For this distribution, $\mu = 1/\lambda$ and integrating by parts we obtain,

$$\sigma_-^2 = \int_0^{1/\lambda}\left(\frac{1}{\lambda} - x\right)^2 \lambda e^{-\lambda x}\,dx = (1-2/e)\sigma^2, \qquad \sigma_+^2 = \int_{1/\lambda}^{\infty}\left(x - \frac{1}{\lambda}\right)^2 \lambda e^{-\lambda x}\,dx = (2/e)\sigma^2$$

For this distribution, the Asymmetry $A = (\sigma_+^2 - \sigma_-^2)/\sigma^2 = 4/e - 1 \approx 0.47$ and Tailness $T = e^2/2 \approx 3.69$. By contrast, the classical Skewness and Kurtosis for exponential distribution are $S = 2$ and $K = 9$ respectively. These values are noticeably larger than the corresponding A and T metrics.

A summary comparison of these distributions is given in Table 1.

Table 1. Comparison of Alternative Distribution Metrics with Classical Distribution Metrics for Several Common Distributions

Distribution	Gaussian	Exponential	Laplace
Mean (μ)	μ	$1/\lambda$	μ
Median (M)	M = μ	$(\log 2)/\lambda$	μ
Standard Deviation	σ	$1/\lambda$	$b\sqrt{2}$
Mean Deviation d	$\sigma\sqrt{2/\pi} \approx 0.80\,\sigma$	$(2/e)\sigma \approx 0.74\,\sigma$	b
σ_-^2	$0.5\sigma^2$	$(1-2/e)\sigma^2 \approx 0.26\,\sigma^2$	b^2
σ_+^2	$0.5\sigma^2$	$(2/e)\sigma^2 \approx 0.74\,\sigma^2$	b^2
Tailness T	$\pi \approx 3.14$	$e^2/2 \approx 3.69$	4
Kurtosis K	3	9	6
Asymmetry A	0	$4/e$ - $1 \approx 0.47$	0
Skewness S	0	2	0

The Alternative Metrics reintroduce the natural constants. In Table 1, the Tailness for the Normal Distribution is π, while the Tailness of the Exponential Distribution is $e^2/2$. Correspondingly, the Asymmetry of the Exponential Distribution is $4/e - 1$.

As expected, the numerical values of $T = 3.69$ for the Exponential and $T = 4$ for Laplace Distributions are noticeably muted compared to the corresponding values of K for these distributions, 9.0 and 6.0, respectively. Note that for the

Exponential distribution Kurtosis $K = 9$ is greater than for the Laplace distribution where $K = 6$. However, for the exponential, Tailness $T = 3.69$ is less than for the Laplace distribution where Tailness $T = 4.0$. The latter inequality as compared to the former Classical inequality might be partially explained because the Laplace distribution has two fat tails while the Exponential Distribution has only one fat tail.

We have made the point that the Variance, σ^2, while treated in Classical Statistics as a single number, is the sum of two numbers, $\sigma^2 = \sigma_-^2 + \sigma_+^2$. This composition of σ^2 is well illustrated in Table 1 for the Case of the asymmetric Exponential Distribution: $\sigma_-^2 = (1 - 2/e)\sigma^2$ and $\sigma_+^2 = (2/e)\sigma^2$. Clearly, the sum of σ_-^2 and $(\sigma)_+^2$ is σ^2. The Asymmetry is then $A = (s_+^2 - s_-^2)/s^2 = 4/e - 1 \approx 0.47$.

3 A Case Study with Empirical Datasets

To illustrate the utility of the Alternative metrics, we consider the following empirical datasets: the NHANES Height, Cholesterol, and Creatinine data and the 2018 Boston Marathon finishing times data [12–14]. These are presented in Table 2.

Table 2. Comparison of Alternative and Classical metrics for Empirical Datasets

Dataset	Cholesterol		Height		Creatinine		Marathon	
metrics	Female	Male	Female	Male	Female	Male	Female	Male
pop. size	3,461	3,277	2,814	2,630	7,120	6,711	11,982	14,675
Min−Max	79–446	76–431	138–189	148–198	0.3–10.7	0.4–13.9	144–446	128–464
Mean μ	182.94	176.68	159.76	173.53	0.98	1.21	245.32	223.02
Median M	179	172	160	173	0.9	1.2	236	212
d	31.57	31.9	5.59	6.14	0.14	0.17	32.4	36.55
St. dev σ	40.58	40.37	7.02	7.67	0.28	0.38	40.95	45.77
σ/d	1.29	1.27	1.26	1.25	2.00	2.24	1.26	1.25
T	3.31	3.20	3.15	3.12	8.26	9.99	3.19	3.14
K	4.68	4.18	3.07	2.95	445.98	390.55	3.40	3.45
A	0.19	0.19	0.04	0.01	0.71	0.76	0.26	0.28
S	0.82	0.75	0.15	0.01	15.82	15.52	0.85	0.91

Table 2 presents results comparing Classical and Alternative distribution shape metrics for the four datasets in a side-by-side fashion to facilitate comparison. The first two datasets are taken from the 2018 National Health and Nutritional Examination Survey [13]. They are the female and male values of Height and Serum Cholesterol. The Serum Creatinine data are from the 1994 NHANES [14]. Creatinine, part of the routine laboratory chemistry profile, is a metabolite of creatine, a muscle component excreted in the urine. Because males

typically have more muscle mass than females, they have higher mean values. The fourth dataset comprises the female and male finishing times of the 2018 Boston Marathon [12].

The following are some of the observations:

- Ratio of Standard Deviation to Mean Deviation, σ/d: we see in Table 2 that for Cholesterol, Height and Marathon, the Standard Deviation is about 25% greater than the Mean Deviation for both males and females. This ratio of about 1.25 is similar to that of the normal distribution and is due to the weight given to the outliers. The Creatinine dataset for both males and females is notable for the very long tails in the empirical PDFs. In the Creatinine data, the Standard Deviation for males, $\sigma = 0.38$, is more than double the Mean Deviation $d = 0.17$ due to the exaggerated influence of these much longer tails. As a result, for both male and female Creatinine, ratios of σ/d are of the order of 2, noticeably higher than the roughly 1.25 ratio for the other three data sets.
- Tailness T and Classical Kurtosis K for the approximately normal male and female Height have comparable values in the general range of 3.0–4.0 as seen in Table 2. By contrast, a very convincing example of the practical utility of Tailness vs. Kurtosis is demonstrated in the case of creatinine.
  We compare Tailness vs. Kurtosis for Creatinine. In Table 2, Classical Kurtosis for the male Creatinine $K = 390.55$ and for the female $K = 445.98$. These numerical results are a typical case where the 4^{th} Moment blows up. In the case of Creatinine, the greatest contribution to the Variance and Kurtosis arises from the outlier points in the very long tail. These highly elevated patient data points likely reflect patients with renal failure and perhaps even dialysis. Yet, these patients are still an intrinsic component of this population-based sample of individuals. There is no scientific reason to prune these high values.
  By contrast with the very high values for Kurtosis, the Tailness for the male Creatinine $T = 9.99$ and for the female $T = 8.26$ are orders of magnitude lower. Even for a noticeably narrowly peaked and long-tailed distribution like the Creatinine, the numerical values for leptokurtic Tailness are a more reasonable multiple of the Tailness $= \pi \approx 3.1416$ of the Normal Distribution.
- Asymmetry vs. Skewness: Suppose that we wish to compare the Asymmetry A for two datasets across two different measurement systems. If the Asymmetry of dataset 1 is $A_1 = 0.3$ and The Asymmetry of dataset 2 is $A_2 = 0.6$; then we can say that dataset 2 is twice as right asymmetric or right skewed as dataset 1. We cannot make a similar statement for the Skewness S metric. In Table 2, Classical Skewness for the female Creatinine $S = 15.82$. Classical Skewness for female Marathon finishing times is $S = 0.85$. If we tried to compare these two measures of Skewness by taking the ratio, we might be tempted to say that the Creatinine is $15.82/0.85 = 18.61$ times more right-skewed than the Marathon finishing times. Such a direct comparison of these values would not make statistical sense. By contrast, the normalized Asymmetry for the female Creatinine is $A = 0.71$ while the normalized Asymmetry for the Marathon

is $A = 0.26$. Then the ratio $0.71/0.26 = 2.73$ allows us to assert that the Creatinine empirical distribution is about 2.73 times more asymmetric or skewed to the right than the Marathon empirical distribution.

The above numerical examples highlight the benefits of using Mean Deviation, Tailness, and Asymmetry to analyze and compare empirical datasets.

4 Summary and Conclusion

We believe the proposed metrics will provide valuable and understandable tools for data analysis. We summarize some of the comparisons between the proposed and classical metrics:

- Simpler to compute and interpret: the Alternative metrics use only 1^{st} and 2^{nd} Moments, whereas classical metrics use higher order moments (for skew and kurtosis). By computing the upside and downside components of variance, we gain additional information to calculate Asymmetry and provide more information about both tails.
- Mean Deviation d is an unbiased measure of data dispersion and does not exaggerate the impact of outliers. By contrast, the Standard Deviation σ is a biased measure of data dispersion that exaggerates the impact of tails and may influence the final result.
- Tailness T that uses the 2^{nd} is more robust to outliers than classical Kurtosis, K that uses the 4^{th} moment. By using Tailness, it may be less necessary to prune outlier points.
- Asymmetry A is always in the range $(-1, 1)$ and therefore allows us to compare datasets across different measurement scales. By contrast, the classical measure of skewness S which is not bounded cannot be used to compare datasets across different measurement scales.

We hope that users will appreciate the additional informational value provided by the new metrics.

References

1. Pham-Gia, T., Hung, T.L.: The mean and median absolute deviations. J. Math. Comput. Modell. **34**, 921–936 (2001)
2. Dodge, Y.: Statistical Data Analysis Based on the L_1 Norm and Related Topics. North-Holland, Amsterdam (1987)
3. Elsayed, K.M.T.: Mean absolute deviation: analysis and applications. Int. J. Bus. Stat. Anal. **2**(2), 63–74 (2015)
4. Gorard, S.: An absolute deviation approach to assessing correlation. Br. J. Educ. Soc. Behav. Sci. **53**(1), 73–81 (2015)
5. Gorard, S.: Introducing the mean deviation "effect" size. Int. J. Res. Method Educ. **38**(2), 105–114 (2015)
6. Rousseeuw, P.J., Croux, C.: Alternatives to the median absolute deviation. J. Am. Stat. Assoc. **88**(424), 1273–1283 (1993)

7. Farebrother, R.W.: The historical development of the l_1 and l_∞ estimation methods. In: Dodge, Y. (ed.) Statistical data Analysis Based on the L_1-Norm and Related Topics, pp. 37–63. North-Holland, Amsterdam (1987)
8. Portnoy, S., Koenker, R.: The gaussian hare and the Laplacian tortoise: computability of square-error versus Abolute-error estimators. Stat. Sci. **2**(88), 279–300 (1997)
9. Pinsky, E., Klawansky, S.: Mad (about median) vs. quantile-based alternatives for classical standard deviation, skewness, and kurtosis. Front. Appl. Math. Stat. **9** (2023). https://doi.org/10.3389/fams.2023.1206537
10. Johnson, N.L., Kotz, S., Balakrishnan, N.: Continuous Univariate Distributions, vol. 1, 2nd edn. Wiley-Interscience, New York (1994)
11. Johnson, N.L., Kotz, S.: Distributions in Statistics. Wiley, New York (1970)
12. Boston marathon archives; Boston athletic association. Boston Athletic Association. http://registration.baa.org
13. NCHS. Nhanes (2017–2018) laboratory data. Centers for Disease Control and Prevention (2018). https://wwwn.cdc.gov/nchs/nhanes
14. NCHS. National center for health statistics; plan and operation of the third national health and nutrition examination survey, 1988–1994, vital health stat, vol. 1, no. 32, pp. 0094–1308. DHHS Publication (1994)