The p-folded Cumulative Distribution Function and the Mean Absolute Deviation from the p-quantile

Jing-Hao Xue[a,*], D. Michael Titterington[b]

[a] *Department of Statistical Science, University College London, London WC1E 6BT, UK*
[b] *School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QQ, UK*

Abstract

The aims of this short note are two-fold. First, it shows that, for a random variable X, the area under the curve of its folded cumulative distribution function equals the mean absolute deviation from the median (MAD). Such an equivalence implies that the MAD is the area between the cumulative distribution function (CDF) of X and that for a degenerate distribution which takes the median as the only value. Secondly, it generalises the folded CDF to a p-folded CDF, and derives the equivalence between the area under the curve of the p-folded CDF and the weighted mean absolute deviation from the p-quantile (MAD_p). In addition, such equivalences give the MAD and MAD_p simple graphical interpretations. Some other practical implications are also briefly discussed.

Keywords: Cumulative distribution function (CDF), Folded CDF, Mean absolute deviation from the median (MAD)

1. Introduction

The folded cumulative distribution function for a random variable can be easily obtained by folding down the upper half of the cumulative distribution function (CDF). It is a simple graphical method for summarising distributions, and has been used for the evaluation of laboratory assays, clinical trials and quality control (Monti, 1995; Krouwer and Monti, 1995).

*Corresponding author. Tel.: +44-20-7679-1863; Fax: +44-20-3108-3105.
Email addresses: `jinghao@stats.ucl.ac.uk` (Jing-Hao Xue), `michael.titterington@gla.ac.uk` (D. Michael Titterington)

The mean absolute deviation from the median (MAD) is obtained by averaging the absolute deviations over a population from its median. It is a summary statistic for measuring the variability or dispersion of a distribution.

This short note first shows that the area under the curve of the folded CDF equals the MAD, and then generalises the folded CDF to a p-folded CDF and derives the equivalence between the area under the curve of the p-folded CDF and the weighted mean absolute deviation from the p-quantile, which has been used as a risk measure for portfolio optimisation (Ogryczak and Ruszczyński, 2002; Ruszczyński and Vanderbei, 2003).

2. Relationship between the folded CDF and the MAD

Consider a univariate, continuous random variable X, with probability density function (PDF) $f(x)$, with CDF $F(x)$ and with the support of $f(x)$ being the interval $[a, b]$. For a discrete X, a derivation similar to the one below can be obtained and is thus omitted here.

2.1. The theoretical case

The CDF $F(x)$ is a real-valued function in the range of $[0, 1]$, defined as

$$F(x) = \int_a^x f(y)dy \ . \tag{1}$$

The folded CDF, denoted by $G(x)$ hereafter, is obtained by folding down the upper half of the CDF. It is therefore a real-valued function in the range of $[0, \frac{1}{2}]$, defined by

$$G(x) = \begin{cases} F(x), & \text{if } F(x) \leq \frac{1}{2} \ , \\ 1 - F(x), & \text{otherwise} \ . \end{cases} \tag{2}$$

A folded CDF is also termed a mountain plot, in view of its shape.

The MAD is defined by

$$\text{MAD} = \int_a^b |x - m| f(x)dx \ , \tag{3}$$

where m is the median of the distribution $F(x)$ such that

$$\int_a^m f(x)dx = \int_m^b f(x)dx = \frac{1}{2} \ . \tag{4}$$

By elementary algebra and interchange of variables for integration, it follows that the area under the curve of $G(x)$ is

$$\begin{aligned}\int_a^b G(x)dx &= \int_a^m F(x)dx + \int_m^b \{1 - F(x)\}dx \\ &= \int_a^m \left\{\int_a^x f(y)dy\right\} dx + \int_m^b \left\{\int_x^b f(y)dy\right\} dx \\ &= \int_a^m \left\{\int_y^m dx\right\} f(y)dy + \int_m^b \left\{\int_m^y dx\right\} f(y)dy \\ &= \int_a^b |y - m| f(y)dy \ . \end{aligned} \quad (5)$$

That is, the area under the curve of $G(x)$ equals the MAD.

2.2. The empirical case

Suppose that we have a sample of N observations from the distribution $F(x)$ and that, among the N observations, there are n distinct values $\{x_i\}_{i=1}^n$ with corresponding proportions $p(x_i)$. Without loss of generality, let $x_1 < x_2 < \ldots < x_n$.

By abuse of notation, we use the same symbols for $F(x)$, $G(x)$, m, MAD and their empirical versions, when there is no ambiguity in the context.

The empirical CDF, $F(x)$, can be defined as

$$F(x) = \sum_{x_i \le x} p(x_i) \ . \quad (6)$$

Empirically, the median m is any point such that

$$F(m) \ge \frac{1}{2} \ \text{ and } \sum_{x_i \ge m} p(x_i) \ge \frac{1}{2} \ . \quad (7)$$

If $m = x_K$ and $m = x_{K+1}$ both satisfy (7) then any x-value such that $x_K \le x \le x_{K+1}$ qualifies to be the sample median. Otherwise, m is the unique x_K for which (7) holds and in this case both inequalities are strict; this argument includes the case in which all the N observations are distinct.

Hence, the area under the curve of $G(x)$ can be expressed as

$$\begin{aligned}
& \sum_{i=1}^{K-1} \{G(x_i)(x_{i+1} - x_i)\} + G(x_K)(m - x_K) \\
+ \quad & G(m)(x_{K+1} - m) + \sum_{i=K+1}^{n-1} \{G(x_i)(x_{i+1} - x_i)\} \\
= \quad & \sum_{i=1}^{K-1} \{F(x_i)(x_{i+1} - x_i)\} + F(x_K)(m - x_K) \\
+ \quad & \{1 - F(m)\}(x_{K+1} - m) + \sum_{i=K+1}^{n-1} [\{1 - F(x_i)\}(x_{i+1} - x_i)] \ . \qquad (8)
\end{aligned}$$

If we substitute equation (6) into equation (8), the area becomes

$$\begin{aligned}
& \sum_{i=1}^{K-1} \left\{ (x_{i+1} - x_i) \sum_{j=1}^{i} p(x_j) \right\} + (m - x_K) \sum_{j=1}^{K} p(x_j) \\
+ \quad & (x_{K+1} - m) \sum_{j=K+1}^{n} p(x_j) + \sum_{i=K+1}^{n-1} \left\{ (x_{i+1} - x_i) \sum_{j=i+1}^{n} p(x_j) \right\} \\
= \quad & \sum_{j=1}^{K} \{(m - x_K + x_K - x_{K-1} + \ldots + x_{j+1} - x_j)p(x_j)\} \\
+ \quad & \sum_{j=K+1}^{n} \{(x_{K+1} - m + x_{K+2} - x_{K+1} + \ldots + x_j - x_{j-1})p(x_j)\} \\
= \quad & \sum_{j=1}^{K} \{(m - x_j)p(x_j)\} + \sum_{j=K+1}^{n} \{(x_j - m)p(x_j)\} \\
= \quad & \sum_{j=1}^{n} \{|x_j - m| p(x_j)\} \ . \qquad (9)
\end{aligned}$$

As the MAD can be defined as

$$\text{MAD} = \sum_{i=1}^{n} \{|x_i - m| p(x_i)\} \ , \qquad (10)$$

equation (9) shows that the area under the curve of $G(x)$ equals the MAD.

Furthermore, equations (5) and (9) suggest that the MAD is the area, or a measure of absolute difference, between $F(x)$ and the CDF for a degenerate distribution which takes the median m as the only value.

3. Generalisations to the p-folded CDF and the MAD_p

The folded CDF can be generalised to a p-folded CDF, denoted by $G_p(x)$ hereafter and given by

$$G_p(x) = \begin{cases} F(x), & \text{if } F(x) \leq p \ , \\ 1 - F(x), & \text{otherwise} \ , \end{cases} \tag{11}$$

where $p \in (0, 1)$.

Similarly, the MAD can also be generalised to a mean absolute deviation from the p-quantile, denoted by MAD_p hereafter and given by

$$\text{MAD}_p = \int_a^b |x - m_p| f(x) dx \ , \tag{12}$$

where, for $p \in (0, 1)$, $m_p = F^{-1}(p)$ is the p-quantile.

Then, as implied by equation (5), the p-folded CDF is related to the MAD_p through $\int_a^b G_p(x)dx = \text{MAD}_p$. In addition, the MAD_p is a measure of absolute difference between $F(x)$ and the CDF for a degenerate distribution which takes m_p as the only value.

However, when p is a value other than 1/2, $G_p(x)$ is not continuous at m_p. Hence, here we define $G_p(x)$ as a weighted version of that in equation (11):

$$G_p(x) = \begin{cases} \frac{1-p}{p} F(x), & \text{if } F(x) \leq p \ , \\ 1 - F(x), & \text{otherwise} \ , \end{cases} \tag{13}$$

for $p \in (0, 1)$, such that $G_p(x)$ is continuous at m_p with $G_p(m_p) = 1 - p$.

Accordingly, the MAD_p is defined as a weighted version of that in equation (12):

$$\text{MAD}_p = \int_a^b \max\left\{ \frac{1-p}{p}(m_p - x), x - m_p \right\} f(x) dx \ , \tag{14}$$

such that

$$\begin{aligned}\int_a^b G_p(x)dx &= \int_a^{m_p} \frac{1-p}{p} F(x)dx + \int_{m_p}^b \{1-F(x)\}dx \\ &= \int_a^{m_p} \frac{1-p}{p}(m_p - y)f(y)dy + \int_{m_p}^b (y-m_p)f(y)dy \\ &= \int_a^b \max\left\{\frac{1-p}{p}(m_p-y), y-m_p\right\} f(y)dy; \ . \end{aligned} \tag{15}$$

that is, the weighted MAD_p equals $\int_a^b G_p(x)dx$, the area under the curve of $G_p(x)$.

From equation (14), we can make the following observations. First, when $p = 1/2$, the MAD_p reverts to the MAD. Secondly, the relative weight received by the values of X larger than m_p is $\frac{p}{1-p}$. When $p > 1/2$, $\frac{p}{1-p} > 1$; hence, the values of X larger than m_p receive a heavier weight than that received by the values smaller than m_p, and the larger the p, the larger the relative weight $\frac{p}{1-p}$. Such a pattern reverses if $p < 1/2$. In both cases, it indicates that, roughly speaking, a deviation from m_p to a more extreme situation receives a heavier weight than a deviation from m_p to a less extreme situation, when the overall variability is summarised by the MAD_p.

Therefore, such an MAD_p can be used as a measure of risk, as adopted in mean-risk models for portfolio optimisation by Ogryczak and Ruszczyński (2002), Ruszczyński and Vanderbei (2003), Miller and Ruszczyński (2008) and Choi and Ruszczyński (2008), for example. These studies have discussed the relationship between the MAP_p and expected shortfall, sometimes termed conditional value at risk, average value at risk or expected tail loss.

4. Implications for practice

Our results have a number of practical implications.

First, analogously to the Bland-Altman difference plot (Altman and Bland, 1983; Bland and Altman, 1986, 1999), which is popular in medical statistics and analytic chemistry, the folded CDF is also a graphical tool for assessing agreement between two assays or methods, often by representing the difference between the two assays by a random variable X. Both plots can be readily understood by the users who may not be statisticians or operations research analysts.

Compared with the Bland-Altman difference plot, the folded CDF stresses more the median and tails of the difference. If the two assays are 'unbiased' with each other (Krouwer and Monti, 1995), the median would be close to zero. If the variability between the two assays is large, the width near the bottom of the folded CDF would be large, analogously to a confidence interval.

Complementary to such a width, the area under the curve of the folded CDF is another measure of the variability between the two assays, roughly through visual inspection or precisely through quantitative computation. Therefore, the equivalence between the under-curve area and the MAD suggests, and provides a theoretical justification of, this measure.

Secondly, the weighted mean absolute deviation from the p-quantile, shown as the MAD_p in equation (14), includes the MAD as a special case and, more importantly, has been adopted as a risk measure in mean-risk models for portfolio optimisation. It is well defined and investigated (Ruszczyński and Vanderbei, 2003). Moreover, it is a very generic measure of dispersion or risk, and can be used in other risk-management practice.

Lastly but importantly, the equivalences give the MAD and MAD_p simple graphical interpretations for practitioners from outside the statistics and operations research communities.

Acknowledgments

This work was partly supported by funding to J.-H.X. from the Internal Visiting Programme, under the EU-funded PASCAL2 Network of Excellence. Thanks to the Associate Editor for the comments.

References

Altman, D. G., Bland, J. M., 1983. Measurement in medicine: the analysis of method comparison studies. Journal of the Royal Statistical Society, Series D (The Statistician) 32 (3), 307–317.

Bland, J. M., Altman, D. G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet 327 (8476), 307–310.

Bland, J. M., Altman, D. G., 1999. Measuring agreement in method comparison studies. Statistical Methods in Medical Research 8 (2), 135–160.

Choi, S., Ruszczyński, A., 2008. A risk-averse newsvendor with law invariant coherent measures of risk. Operations Research Letters 36 (1), 77–82.

Krouwer, J. S., Monti, K. L., 1995. A simple, graphical method to evaluate laboratory assays. European Journal of Clinical Chemistry and Clinical Biochemistry 33 (8), 525–527.

Miller, N., Ruszczyński, A., 2008. Risk-adjusted probability measures in portfolio optimization with coherent measures of risk. European Journal of Operational Research 191 (1), 193–206.

Monti, K. L., 1995. Folded empirical distribution function curves–mountain plots. The American Statistician 49 (4), 342–345.

Ogryczak, W., Ruszczyński, A., 2002. Dual stochastic dominance and related mean-risk models. SIAM Journal on Optimization 13 (1), 60–78.

Ruszczyński, A., Vanderbei, R. J., 2003. Frontiers of stochastically nondominated portfolios. Econometrica 71 (4), 1287–1297.