

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

Jing-Hao Xue<sup>a,\*</sup>, D. Michael Titterington<sup>b</sup>

<sup>a</sup>*Department of Statistical Science, University College London, London WC1E 6BT, UK*

<sup>b</sup>*School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QQ, UK*

---

## Abstract

The aims of this short note are twofold. First, it shows that, for a random variable  $X$ , the area under the curve of its folded cumulative distribution function (or the mountain plot) equals the mean absolute deviation from the median (MAD). Such an equivalence implies that the MAD is the area (or a measure of absolute difference) 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

2 The folded cumulative distribution function for a random variable, also  
3 termed the mountain plot, can be easily obtained by folding down the upper  
4 half of the cumulative distribution function (CDF). It is a simple graphical  
5 method for summarising distributions, and has been used for the evaluation  
6 of laboratory assays, clinical trials and quality control (??).

---

\*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)

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

11 This short note first shows that the area under the curve of the folded  
12 CDF equals the MAD, and then generalises the folded CDF to a  $p$ -folded  
13 CDF and derives the equivalence between the area under the curve of the  $p$ -  
14 folded CDF and the (weighted) mean absolute deviation from the  $p$ -quantile,  
15 which has been used as a risk measure for portfolio optimisation (??).

## 16 2. Equivalence between the folded CDF and the MAD

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

### 21 2.1. The theoretical case

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

$$F(x) = \int_a^x f(y)dy . \quad (1)$$

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

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

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

27 The MAD is defined by

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

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

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

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

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

31 That is,  $\int_a^b G(x)dx = \text{MAD}$  .

### 32 2.2. The empirical case

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

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

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

$$F(x) = \sum_{x_i \leq x} p(x_i) . \tag{6}$$

40 Empirically, the median  $m$  is any point such that

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

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

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

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

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

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

47 As the MAD can be defined as

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

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

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

### 52 3. Generalisations to $p$ -folded CDF and $MAD_p$

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

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

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

56 Similarly, the MAD can also be generalised to a mean absolute deviation  
 57 from the  $p$ -quantile (denoted by  $MAD_p$  hereafter), given by

$$MAD_p = \int_a^b |x - m_p| f(x) dx, \quad (12)$$

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

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

63 However, when  $p$  is a value other than  $1/2$ ,  $G_p(x)$  is not continuous at  
 64  $m_p$ . Hence, here we define  $G_p(x)$  as a weighted version of that in equation  
 65 (11):

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

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

67 Accordingly, the  $MAD_p$  is defined as a weighted version of that in equation  
 68 (12):

$$MAD_p = \int_a^b \max \left\{ \frac{1-p}{p} (m_p - x), x - m_p \right\} f(x) dx, \quad (14)$$

69 such that

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

70 That is, the (weighted)  $\text{MAD}_p$  equals  $\int_a^b G_p(x)dx$ , the area under the curve  
 71 of  $G_p(x)$ .

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

81 Therefore, such an  $\text{MAD}_p$  can be used as a measure of risk, as adopted  
 82 in mean-risk models for portfolio optimisation by ?, ?, ? and ?, for example.  
 83 In these studies, the relationship between the  $\text{MAP}_p$  and expected shortfall  
 84 (sometimes termed conditional value at risk, average value at risk or expected  
 85 tail loss) has also been discussed.

#### 86 4. Implications for practice

87 Our results have a number of practical implications.

88 First, analogously to the Bland-Altman difference plot (???), which is  
 89 hugely popular in medical statistics and analytic chemistry, the folded CDF  
 90 is also a graphical tool for assessing agreement between two assays (or meth-  
 91 ods), often by representing the difference between the two assays by a random  
 92 variable  $X$ . Both plots can be readily understood by the users who may not  
 93 be statisticians or operational research analysts.

94 Compared with the difference plot, the folded CDF stresses more the  
 95 median and tails of the difference. If the two assays are ‘unbiased’ with each

96 other (?), the median would be close to zero. If the variability between the  
97 two assays is small, the width near the bottom of the folded CDF would be  
98 large (?), analogously to a confidence interval.

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

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

110 Lastly but importantly, the equivalences give the MAD and  $MAD_p$  sim-  
111 ple graphical interpretations for practitioners from outside the statistics and  
112 operational research communities.

### 113 **Acknowledgments**

114 This work was partly supported by funding to J.-H.X. from the Internal  
115 Visiting Programme, under the EU-funded PASCAL2 Network of Excellence.