管理经典

经济发展与价值选择

陈春花◎著

Economic Development and the Value Choice

陈春花管理思想的形成源于哲学思考

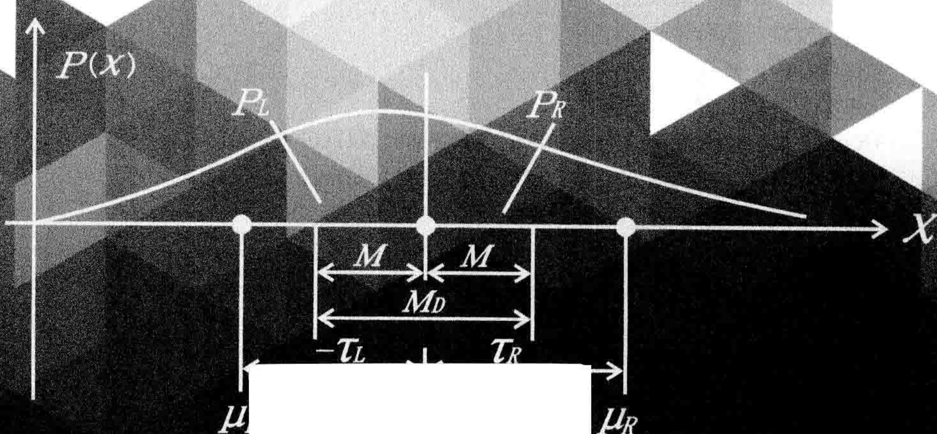
企业存在的意义在于为顾客创造价值，人存在的意义在于自我实现的价值

机械工业出版社
China Machine Press

概率统计中的平衡定理及线性特征参数

BALANCE THEOREM FOR PROBABITY STATISTICS
AND
LINEAR CHARACTERISTIC PARAMETERS

黄志新 著



中山大學出版社
SUN YAT-SEN UNIVERSITY PRESS
·广州·

图书在版编目（CIP）数据

概率统计中的平衡定理及线性特征参数/黄志新著.—广州：中山大学出版社，2016.8

ISBN 978-7-306-05725-9

Ⅰ.①概… Ⅱ.①黄… Ⅲ.①概率统计—研究 Ⅳ.①O211

中国版本图书馆 CIP 数据核字（2016）第 135093 号

出 版 人：徐 劲
策划编辑：李 文
责任编辑：黄浩佳
封面设计：曾 斌
责任校对：李艳清
责任技编：何雅涛
出版发行：中山大学出版社
电 话：编辑部 020-84111996，84113349，84111997，84110779
发行部 020-84111998，84111981，84111160
地 址：广州市新港西路 135 号
邮 编：510275 传 真：020-84036565
网 址：http://www.zsup.com.cn E-mail：zdcbs@mail.sysu.edu.cn
印 刷 者：虎彩印艺股份有限公司
规 格：787mm×1092mm 1/16 14.75 印张 347 千字
版次印次：2016 年 8 月第 1 版 2016 年 8 月第 1 次印刷
定 价：40.00 元

如发现本书因印装质量影响阅读，请与出版社发行部联系调换

作 者 简 介

黄志新，广州市荔湾区科协成员。主要研究方向为“概率统计的线性特征参数及其应用”。现正把该项目扩展至贝叶斯统计上去，用以克服先验分布难求这一瓶颈。曾经发表论文多篇，主要有《概率统计的平衡定理》《哈勃红移的量子理论解释》《人造太阳》《基于“东方之星”的翻沉，看旅游船的稳性缺陷及对策》等等。参与开发中国专利“防近视台灯”、防止二次供水污染的“水池密封装置”、“自行车安全升降器”等专利。

内 容 简 介

概率统计数据包含丰富的信息资源。对于统计量的诸多信息，现在常用数字特征以均值、方差（或标准差）来表达，但这远远不能表达统计数据包含的丰富内涵，实为挂一漏万！特别是反映随机变量的离散程度，现在通常用方差（或标准差）来表征。这两个特征参数有以下的缺点：

(1) 只能一般地反映数据偏离均值的情况，而不能反映是正偏或负偏等情况。

(2) 样本的标准差比方差更能反映数据的离散情况，但样本标准差是其总体标准差的有偏估计，所以少用。

(3) 现在通常用样本修正方差反映数据的离散情况，但方差的单位是均值的平方，这既不合理也不科学。

(4) 样本的方差的稳定性是最差的，只要有一些干扰，就会被“平方”地放大了，变得面目全非。

(5) 统计数据包含了太多的原始信息资料，单凭均值、标准差（方差）远远不能反映这些信息：如偏态情况（正偏、负偏及无偏），峰态情况等等，则需用更难计算的三阶矩、四阶矩才能表达。而三阶矩及四阶矩的计算既繁琐又存在着其他很多缺点。

作者二十年磨一剑，在概率统计领域依据一条最普通的定理“一阶中心矩恒为0”进行开创性的研究工作，由此出发推导出了“概率统计的平衡定理”以及“三均值公式”，提出了一系列与以往不同、全新的特征参数：右均值和左均值，正均差和负均差，右概率和左概率，以及半均差，重新研究了平均差。研究表明，这些特征参数种类繁多，从各个不同的侧面反映了统计数据所含的信息，它们都是统计数据的线性函数，由于线性函数具有良好的性质，因此基本上克服了标准差与方差上面所述的缺点。在概率统计领域开辟了一片新天地——线性特征参数统计，这一统计方法与现在的参数统计、非参数统计及贝叶斯统计并驾齐驱，它们各有各的特点，各有各的用途。特征参数统计既具有“参数统计”精确度高的优点，又具有不需要数据一定为“正态分布”，使用广泛而简便的优点，但是它比“非参数统计”精确得多。因此，可以预料它必将成为一种新的有用的统计方法。

作者现在把这一理论充分开拓、壮大、独成一派，并反复修改提炼，写成

一本精品专著:《概率统计中的平衡定理及线性特征参数》，贡献给读者。本书是在一篇已发表论文的基础上创作而成，基本上是原创的。除了保留平均值(数学期望)，及个别地方之外，其他的概念、定义、定理、公式与现有的理论很少共通之处，在这点上，希望读者注意。著名数学家G. 康托尔在1883年曾经说过，数学在它自身的发展中完全是自由的，对它的概念的限制只在于：必须是无矛盾性的，并且和先前由确切定义引进的概念相协调的。请读者用这一观点衡量本书。

现在恰逢党中央、国务院大力提倡创造、创新精神。“线性特征参数统计”是由中国人首创的，顺应了这一伟大的潮流。它是一片富饶、美丽的科学新大陆。它可以在概率统计现有的领域广泛应用：如数学专业（特别是概率统计专业)、理工科、工程实践、经济财务、科学实验等等。

欣赏这片富饶、美丽的科学新大陆，读者只需一般大学水平就可以了。但是，它还有大量未被发掘的宝贵矿藏，如：特征参数在贝叶斯统计中的应用等等，具有无限的吸引力，定将吸引一大批有志于创新的本科生、硕士生、博士生以及科研人员前来开采。若有意共同合作研究，或购买此书者，请电邮871367655@qq.com，或发微信、短信：13005107434，13710639807，与作者联系，共同研究和开发这块尚未完全开垦的处女地。

前　言

概率统计数据包含的丰富信息，只凭均值和方差（或均方差）是远远不能表达其所有内涵的；而三阶矩及四阶矩计算既繁琐又存在着很多缺点.

概率统计中，用方差或均方差来反映随机变量的离散程度，这两个特征参数有以下缺点：样本修正方差 $\tilde{s}_{n-1}^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ 是总体方差 σ^2 的无偏估计；样本方差 $s_n^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ 是 σ^2 的极大似然估计；$\tilde{s}_{n+1}^2=\frac{1}{n+1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ 是以上三者中对 σ^2 的极小均方误差估计，及最优同变估计；但后两者并不是 σ^2 的无偏估计，见表 1.

表 1

符号	$\tilde{s}_{n-1}^2$	$\tilde{s}_n^2$	$\tilde{s}_{n+1}^2$
表达式	$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$	$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$	$\frac{1}{n+1}\sum_{i=1}^{n}(x_i-\bar{x})^2$
对总体 σ^2 的估计	最小方差无偏估计	矩估计 极大似然估计	极小均方误差估计 及最佳同变估计
用途	无偏估计，常用	有偏估计，有时用	有偏估计，少用

虽然 $\tilde{s}_{n-1}^2$ 是 σ^2 的无偏估计，但是决不能推断标准差 $\tilde{s}_{n-1}$ 就是 σ 的无偏估计！σ 的最小方差无偏估为：$\tilde{\sigma}=k\tilde{s}_{n-1}=k\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x}^2)}$，变系数 k 是 Γ 积分的复杂的函数，它会随着样本数量 n 的变化而变化. 理论上 σ 比 σ^2 更能反映随机变量的离散程度，因为 σ 的量纲与均值的量纲一致，而 σ^2 的量纲与均值却不同，由于这一系列的矛盾，反映了方差和标准差并不是随机变量的离散程度的最佳表征.

对于统计量的诸多信息，仅靠方差（或标准差）是反映不全的，如随机变量的偏态情况：正偏、负偏、无偏及峰态等需用更难计算的三、四阶矩才能

表达.

概率统计中，有一条最平常，而又最被人忽视的定理：一阶中心矩恒为0. 换种思维方式，本文定义了“半均差”，计算半均差不用绝对值，而且计算过程很容易；而平均差就等于两倍的半均差. 结合概率统计中上述这一条众所周知的定理，本文推导出一条全新的“平衡定理”，以及三均值公式，由此出发推导出它一系列与以往不同、全新的数学特征参数：右均值和左均值，正均差和负均差，右概率和左概率，重新研究了平均差. 研究表明，这些特征参数都是统计量的线性参数，因而被称为“线性特征参数”，并且具有表征随机变量数据集的以下功能：反映数据的离散程度（具有均方差及方差的功能）；反映对均值的偏态程度（具有偏态系数的功能）；反映峰态程度，（具有峰态系数的功能）；样本的线性特征参数对总体同名参数是无偏估计；数学意义明确；信息多. 平衡定理是反比例方程，利用它很容易计算上述那几个新的特征参数. 这些新特征参数既有 σ 及 σ^2 的功能，但却没有它们的缺点，而且其所含的信息量也比 σ 及 σ^2 大得多.

现在所有的特征参数只是描述（或估计）其数据的总体（或样本）的情况，除分位数之外，基本上没有描述局部的特征参数. 本文的第 6 章是专门研究描写随机变量各部分的一些特征参数，称为各级线性特征参数，这需要与各阶特征参数相区别.

第 8 章是讲多维随机变量的线性特征参数，这里提出了一些与现有方案不同的方法，有着非常丰富的内容，谨供大家参考.

因为是线性特征参数，在样本对总体的点估计中，它们都有着共同的优点：样本某一线性特征参数都是对其总体同名特征参数的一致最小方差最优无偏估计，并且对于正态分布来说，是它们的极大似然估计以及最优线性同变估计. 我国著名的数理统计学家陈希孺教授曾经说过：“好的点估计会产生好的区间估计.”除此之外，还会产生其他好的统计效果. 因此，以上的两条公式定理及由此而产生的各种新的数学特征参数在概率统计及其他学科中有广泛的应用. 这是第 9 章要阐述的内容.

第 10 章是谈线性特征参数在相关分析中的应用（相当于现有的回归分析）.

现有的统计学包括几个相互独立的不同的分支或派系：①参数统计. ②非参数统计. ③贝叶斯统计. 本书所讲内容不属于上述统计中的任何一种，而是新开拓的一种统计学，被称为“线性特征参数统计学”. 它介于参数统计与非参数统计之间，既有参数统计较为精确的优点，又有非参数统计不限于正态分布而适应广泛等优点，还不需要知道数据为何种分布.

著名数学家 G. Contor 在 1883 年曾经说过，数学在它自身的发展中完全是自由的，对它的概念的限制只在于：必须是无矛盾性的，并且是和先前由确切定义引进的概念相协调的.

本书是在一篇已发表的论文的基础上创作而成的，基本上具有原创性. 除了数学期望（均值）及个别地方之外，其他的概念、定义、定理、公式与现有的理论很少有共通之处，在这点上，希望读者注意，请按 G. Contor 所说的无矛盾性与相协调性来衡量本书. 由于本人水平有限，错漏在所难免，诚恳希望读者提出指正，并一起进行深入的研究.

本书适合本科高年级学生、研究生、概率统计应用人员及有志于研究的学者阅读. 第 1 ～ 7 章的阅读只要有高等数学及概率统计知识的基础便可；第 8 章的阅读需要线性代数知识的基础；第 9、10 章的阅读需要参数及非参数统计的知识的基础；第 11 章的阅读要有线性回归知识的基础.

本书的出版得到很多有识之士的支持与帮助，特别是中山大学的王向阳教授，他细心地审阅了本文的第 1 ～ 3 章，并对错漏之处作了修改；澳门科技大学原校长著名教授许敖敖组织了一些教授阅读了本书的一些章节，并给予基本的肯定. 此外还得到中山大学出版社的李文主任、黄浩佳编辑的大力支持；黄俊强先生绘制了本书所有的插图；以及其他很多人的帮助. 没有他们的支持与帮助，本书的出版几乎是不可能的，在此，笔者一并向他们致谢！

关键词：平衡定理与三均值公式；右均值与左均值；正均差与负均差；右概率与左概率；半均差与平均差.

本书所用新的数学名称解释：

1. 特征参数：即相当于一般统计书籍所写的“数字特征”，但前者比后者所包括的范围广得多，正由于此，本书特正名为特征参数，包括数学期望、方差、均方差、三阶矩、四阶矩及以下所述的各种线性特征参数. 应注意特征参数与特征函数以及线性代数中的特征值相区别，这三者表示的是三种不同的概念.

2. 线性特征参数：本书新定义的数学名称，只选特征参数中属于线性的部分参数，包括平均值、右均值、左均值；正均差、负均差；右概率和左概率；半均差和平均差；还包括二级以上（含二级）的线性的特征参数. 但不包括方差、均方差，以及其他二阶以上的中心矩.

3. 同名点估计：现在的点估计，一般是用样本的特征参数去估计总体的原有参数，如以样本的均值及方差去估计泊松分布的参数 λ 等等. 而一般的概率参数的意义千奇百怪，各不相同，这就是为什么现在的参数估计是这么麻烦的原因. 同名点估计也是本书新定义的一个新数学名词，它是以样本的特征

参数去估计总体的同名特征参数，如以样本的平均值、右均值、左均值；正均差、负均差；右频率和左频率；半均差和平均差等去估计上述第 2 点所述总体同名的新的线性特征参数．现在统计学界唯一使用并仅有的同名估计为正态分布中使用的，以样本的均值估计总体的期望值；以样本的方差去估计总体的方差 σ^2（就是同名参数估计）．由于它们的意义明确，使用又方便，估计又准确，所以得到很好的应用．而对于其他分布就都不是同名参数估计了，因此，现在的统计分析，除了正态分布之外，其他分布都不太理想．所以本书对弥补现在统计学的遗憾起到重要作用．

BALANCE THEOREM FOR PROBABILITY STATISTICS AND LINEAR CHARACTERISTIC PARAMETER

Zhi-xin Huang

Abstract: The probability statistical data contains a wealth of information, only by the average and variance (or standard deviation) couldn't express all of their connotation. And that third-order moment and fourth-order moment calculation are very cumbersome and there are a lot of shortcomings. This paper defines some new mathematical characteristic parameters. And derive the balance theorem and the three averages formula. These new characteristic parameter have been following functions: Reflect the degree of dispersion of the data, degree of skewness, the kurtosis. These new sample characteristic parameters of statistics is unbiased estimator of the same name overall characteristic parameters. They also contain a lot of information of overall. So these two formulas and the new characteristics parameters wide range of application in probability and statistics and other disciplines.

Key words: balance theorem; semi-average deviation; positive and negative average deviation; left and right average value; left and right probability

0. Introduction

First central moments constant to 0, that is theorem in the probability and statistics which is the most common and most neglected. [1], ⋯, [4]. Let's exchange another way thinking, a new balance theorem can be deduced, from where proceeding to derive out of it a new series of feature parameters, and their feature is studied.

1. Left and right probability

Definition 1: Let X be the random variable in any probability space $(\Omega, F\ P)$ There is a mathematical expectation μ. Let $P_\mu = P\ \{X = \mu\}$ be the probability of μ.

$$\begin{cases} P_L = P\{X < \mu\} + \dfrac{1}{2}P_\mu \\ P_R = P\{X > \mu\} + \dfrac{1}{2}P_\mu \end{cases} \tag{1}$$

P_L and P_R are the left and right probabilities of X, respectively, where $P\{-\infty < X < +\infty\} = P_L + P_R = 1$.

2. Left and right means and positive and negative mean deviations

1) The expectation of X exists; therefore, the expectations of the left and right distributions exist.

Definition 2: Let the following integrals be in absolute convergence.

$$\begin{cases} \mu_L = E_L(X) = \dfrac{1}{P_L}\left[\dfrac{\mu P_\mu}{2} + \displaystyle\int_{x<\mu} x\mathrm{d}F(x)\right] \\ \mu_R = E_R(X) = \dfrac{1}{P_R}\left[\dfrac{\mu P_\mu}{2} + \displaystyle\int_{x>\mu} x\mathrm{d}F(x)\right] \end{cases} \tag{2}$$

μ_L, μ_R are known as the left and right means of X, respectively, and $\mu_L \leqslant \mu \leqslant \mu_R$.

In oceanography, μ is known as the mean tidal level, μ_L as the mean low tide level and μ_R as the mean high tide level (You et al., 2010).

Definition 3: Let

$$\begin{cases} \tau_L = -\dfrac{1}{P_L}\displaystyle\int_{x \lesssim \mu} (x - \mu)\mathrm{d}F(x) \\ \tau_R = \dfrac{1}{P_R}\displaystyle\int_{x \gtrsim \mu} (x - \mu)\mathrm{d}F(x) \end{cases} \tag{3}$$

$(-\tau_L)$ and (τ_R) are the negative and positive mean deviations of x, respectively.

Equations (1) and (2) are substituted into equation (3). The result is as follows:

$$\begin{cases} -\tau_L = \mu_L - \mu \\ \tau_R = \mu_R - \mu \end{cases}. \tag{4}$$

As the first-order central moment is 0,

$$\int_{-\infty}^{\infty} (x - \mu)\mathrm{d}F(x) = 0,$$

such that

$$\int_{x<\mu} (x - \mu)\mathrm{d}F(x) + \int_{x=\mu} (x - \mu)\mathrm{d}F(x) + \int_{x>\mu} (x - \mu)\mathrm{d}F(x) = 0,$$

as

$$\int_{x=\mu}(x-\mu)\,\mathrm{d}F(x) = 0$$

so that

$$-\int_{x<\mu}(x-\mu)\,\mathrm{d}F(x) = \int_{x>\mu}(x-\mu)\,\mathrm{d}F(x). \tag{5}$$

Definition 4: Let

$$M = -\int_{x<\mu}(x-\mu)\,\mathrm{d}F(x) = \int_{x>\mu}(x-\mu)\,\mathrm{d}F(x)\ . \tag{6}$$

M is the semi-average deviation, also known as the half central moment. From the preceding formula, we can obtain $M \geqslant 0$.

Theorem 1: If the mathematical expectation μ of the random variable X exists, then the average deviation M_D is equal to double the semi-average deviation Ml:

$$M_D = 2M. \tag{7}$$

Proof: According to the mean deviation definition and equation (6), we have:

$$\begin{aligned}
M_D &= \int_{-\infty}^{\infty}|x-\mu|\,\mathrm{d}F(x)\\
&= \int_{x<\mu}|x-\mu|\,\mathrm{d}F(x) + \int_{x=\mu}|x-\mu|\,\mathrm{d}F(x) + \int_{x>\mu}|x-\mu|\,\mathrm{d}F(x)\\
&= -\int_{x<\mu}(x-\mu)\,\mathrm{d}F(x) + \int_{x>\mu}(x-\mu)\,\mathrm{d}F(x)\\
&= 2\int_{x>\mu}(x-\mu)\,\mathrm{d}F(x) = 2M.
\end{aligned}$$

According to this theorem, we know that when we consider μ as the fiducial center point, M_D and are equipollent, i. e. , two different ways of expressing the same numerical characteristics. This is also why M is referred to as the semi-average deviation.

Theorem 2 (Equilibrium Theorem): If the mathematical expectation μ of random variable X exists, then

$$P_L\tau_L = P_R\tau_R = M = M_D/2. \tag{8}$$

Proof: From equations (3) and (6), we obtain:

$$\begin{aligned}
P_L\tau_L &= -P_L\left[\frac{1}{P_L}\int_{x \lessgtr \mu}(x-\mu)\,\mathrm{d}F_L(x)\right]\\
&= -\int_{x \lessgtr \mu}(x-\mu)\,\mathrm{d}F(x) = M\ .
\end{aligned}$$

We can similarly prove that $M = P_R\tau_R$, completing the proof.

Judging from the preceding proof process, M is an invariant from which we can deduce many novel theorems and formulae.

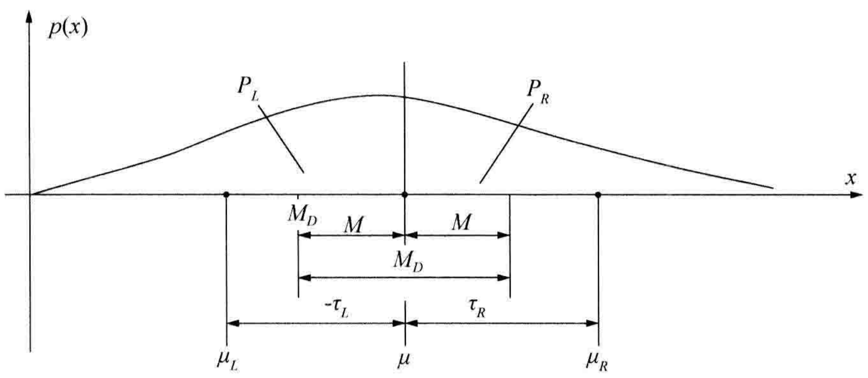


Fig. 1

Deduction 1: Let X be a random variable with a mean μ , a left mean μ_L , a right mean μ_R , a left probability P_L and a right probability P_R . We have the following three-mean value formula:

$$\mu_L P_L + \mu_R P_R = \mu. \tag{9}$$

Proof: Equation (3) is substituted in the first half of equation (8).

$$(\mu - \mu_L)P_L = (\mu_R - \mu)P_R,$$

$$\mu_L P_L + \mu_R P_R = (P_L + P_R)\mu.$$

Equation (9) is proved based on $(P_L + P_R) = 1$. The physical significance of the equilibrium theorem is that it takes the mean value (centroid) as the boundary and divides the value of the random variable (substance) into two. The mathematical significance of μ_L, μ_R lies in the data center or geometrical center of the left and right parts, and their physical significance represents the centroid of the two parts, respectively. $P_L + P_R$ indicate the magnitude of each relevant resultant force, and τ_L and τ_R indicate the average length of each force arm from the resultant force of the left and right parts to the centroid. Equation (8) indicates the static equilibrium. The moments from the left and right sides to the centroid (semi-average deviation) are equal.

Conclusion These characteristic parameters are belong to their overall distribution function of linear functional. They have good linearity and have wide range of uses.

1) Because the number of characteristic parameters is greatly increased, on which can be reflected many aspects characteristic information of the statistical.

2) As a result of these new feature parameters are linear function, so the sample characteristic parameters are their overall respective corresponding parameters of the minimum variance unbiased estimator.

3) Many reference literature reflects the degree deviation of statistics data, which use the mean deviation better than the variance. However, calculate the mean difference must to use absolute value, inconvenient to use. Because of the (10) formula, so to calculate the mean deviation is very easy.

4) The theorem described in this paper and several characteristic parameters, which can be recomposition a new statistics, it is called the Linear Characteristic Parameter Statistics . This statistics is different from the Parametric statistics, Nonparametric statistics, Bayesian statistics. These matter will be detail discourse after this paper.

本文特有的符号

（按文中出现的次序排列）

符号	中文名称	English
$\lim\limits_{x\to xk}(x_k)=\frac{1}{2}[\lim\limits_{x\to x-0}(x_k)+\lim\limits_{x\to x+0}(x_k)]$	中极限	middle limit
$\hat{<}$	小于加上半等于	less and semi-equal to
$\hat{>}$	大于加上半等于	greater and semi-equal to
$P\{X \hat{<} x_k\}$	中连续	middle continuation
$P_L(X)=P\{X \hat{<} \mu\}$	左概率	left probability
$P_R(X)=P\{X \hat{>} \mu\}$	右概率	right probability
$F_L(x)$	左分布函数	left distribution function
$F_R(x)$	右分布函数	right distribution function
$\mu_L = E_L(x)$	左均值	left average
$\mu_R = E_R(x)$	右均值	right average
$-\tau_L = \mu_L - \mu$	负均差	negative mean deviation
$\tau_R = \mu_R - \mu$	正均差	positive mean deviation
M	半均差 （半中心矩）	semi-mean deviation (half central moment)
M_D	绝对平均差	absolute mean deviation
$C_V = M_D/\|\mu\|$	中性离散系数	ordinary dispersed coefficient
$C_{RV} = \tau_R/\|\mu\|$	正离散系数	positive dispersed coefficient
$C_{LV} = \tau_L/\|\mu\|$	负离散系数	negative dispersed coefficient
$S_K = P_L - P_R$	线性偏态系数	linear coefficient of skew
$C_K = P\{\mu_L \hat{<} x \hat{<} \mu_R\}$	线性峰态系数	linear coefficient of kurtosis
P_{2Lj}	二级左区间概率	
P_{2Rj}	二级右区间概率	
μ_{2Lj}	二级区间左均值	
μ_{2Rj}	二级区间右均值	
P_{kLj}	k 级左 j 区间概率	
P_{kRj}	k 级右 j 区间概率	
μ_{kLj}	k 级左 j 区间均值	

μ_{kRj}	k 级右 j 区间均值
$\bar{p}_{kLj}$	k 级左区间的平均概率
$\bar{p}_{kRj}$	k 级右区间的平均概率
$-\tau_{2Lj}$	二级左 j 区间对中心的负均差
τ_{2Rj}	二级右 j 区间对中心的负均差
$-\tau_{kLj}$	k 级左 j 区间对中心的负均差
τ_{kRj}	k 级右 j 区间对中心的负均差
M_{kLij}	k 级左局部半中心矩
M_{kRij}	k 级右局部半中心矩