期望差提出的统计背景及其意义

孔璐[1] 孔粤[2]孔建新[3]

1 广东省中山火炬开发区职业技术学院　中山市（528437）

E-mail：yudianlulu@163.com

2 云南省民族大学继续教育学院 云南省昆明市（650031）

E-mail：baobao_723@yahoo.com.cn

3 云南会泽者海镇 2-24-6 统计科研小组（654211）

E-mail：kongfanjx@163.com

摘　要：在单峰分布的条件下期望差是随机变量与期望值离散程度的变异指标。由于单峰分布包括了对称的正态分布与不对称的偏斜分布。然而传统意义上的标准差是随机变量与平均值离散程度的变异指标。当呈现偏斜分布时分布曲线拐点的确定与期望差相关而与传统意义上的标准差无关。这是提出期望差的统计背景。其意义是：确定任意单峰分布曲线的拐点，为计算其任意期间的概率奠定理论基础。

关键词：单峰分布，偏斜分布，期望差，左、右期望差，拐点

中图分类号：O212.2

0. 引言

单峰分布包括对称的正态分布和不对称的偏斜分布[1]。其数学期望值同样都是分布峰值（众数）点上的取值。在正态分布的条件下平均值与期望值等价，在偏斜分布的条件下平均值与期望值不等。传统意义上的标准差是随机变量与平均值离散程度的变异指标，当呈现正态分布时，平均值加减标准差是其分布曲线的拐点，两拐点间的概率是 0.6826。当呈现不对称的偏斜分布时，平均值加减标准差就不是分布曲线的拐点，两点间的概率就不可能是 0.6826。由于在单峰的条件下还客观存在不对称的偏斜分布[2]，其不对称性决定了以期望值为中线分布曲线两边的离散程度的变异指标也就不相等，分布曲线拐点的确定需要有新的计算方法，在此统计背景下，依据高斯分布的原理并与传统意义上的标准差相区别，根据实际含义，即：随机变量与期望值离散程度的变异指标，相应提出期望差、左期望差、右期望差[3] 的新概念。

提出期望差的意义是：在任意的单峰分布中，分布的期望值减左期望差，加右期望差所对应的分布函数的点是分布曲线的拐点，两拐点间的概率是 0.6826。从而解决了在单峰的条件下确定不对称分布曲线拐点的计算问题。以便应用于统计分析与统计推断的实践中。

本文从统计实践出发再次重新提出期望差的概念，目的是：更深入的讨论提出期望差的统计背景，以便加深对期望差的理解，使之更加广泛的为人们所接受。并从实际含义中进行解读，以便于区别于传统意义上的标准差。同时从实际应用中谈其意义。

1. 期望差提出的统计背景

在统计实践中，随机变量的频数分布在单峰的条件下客观存在对称的正态分布和不对称的偏斜分布这两种形态。在统计分析产品质量指标分布状况时，不同质量特征的产品质量指标所服从的分布显然是不同的。

单峰分布密度函数的建立归纳了以上两种分布形态，由于不对称的客观存在就必然引出了左边随机变量与期望值离散程度变异指标，右边随机变量与期望值离散程度变异指标。为

了与传统标准差的概念相区别，将前者的变异指标称为左期望差，后者的变异指标称为右期望差。使偏斜分布期望值减左期望差，期望值加右期望差所对应分布函数的点成为分布曲线上的拐点。两拐点间的概率就是 0.6826。见图 1 所示。

以下从不同质量特征产品的质量指标所服从分布的实际情况来加以阐明。

在机械加工行业，产品的质量指标围绕目标值波动，目标值是标准规定的最好水平值，且质量控制属于双侧规范，产品质量指标服从正态分布。一般说来平均值等于或接近目标值。平均值加减传统意义上的标准差是其分布曲线上的拐点，两拐点间的概率是 0.6826。但是，当平均值不等于分布的期望值（众数）时，显然出现不对称，分布呈现偏斜，此时平均值加减传统意义上的标准差就不是分布曲线上的拐点，两点间的概率就不是 0.6826。分布曲线拐点的计算需要应用左、右期望差来加以确定。

在冶金行业，特别是食品加工行业的产品，质量控制属于单侧规范，有效成分不能低，需要控制下限；杂质和有害元素不能高，需要控制上限。产品质量指标服从偏斜分布。标准规定的规范限是产品质量的最低水平。最好水平值（目标值）在规范限的另外一端。质量指标围绕分布的期望值（众数）波动，而期望值是无法事先确定的。期望值接近目标值说明质量好，反之则不好。在偏斜分布的情况下，平均值不等于众数，平均值加减传统意义上的标准差不是分布曲线上的拐点，两点间的概率也就不可能是 0.6826。同理，分布曲线拐点的计算同样需要应用左、右期望差来加以确定。

通过统计实践说明左右期望差提出的统计背景概括如下：

在偏斜分布的情况下，期望值不等于平均值，呈现出不对称，以期望值为中线两边随机变量与期望值离散程度的变异指标必然不等，在此统计背景下提出左期望差与右期望差。此时期望差与传统意义上的标准差的含义就完全不同了。前者分别是小于和大于期望值随机变量与期望值离散程度的变异指标；后者是随机变量与平均值离散程度的变异指标。由于统计分析所针对的对象其分布已经发生了变化，呈偏斜分布。若仍然用传统意义上的标准差来描述必然使分析的结果产生很大的偏差。因为传统意义上的标准差是无法确定偏斜分布曲线拐点的。这是提出左右期望差的必要性。提出左右期望差的目的是：在单峰分布的条件下，当呈现不对称的偏斜分布时以确定其分布曲线的拐点。见图 1 所示。

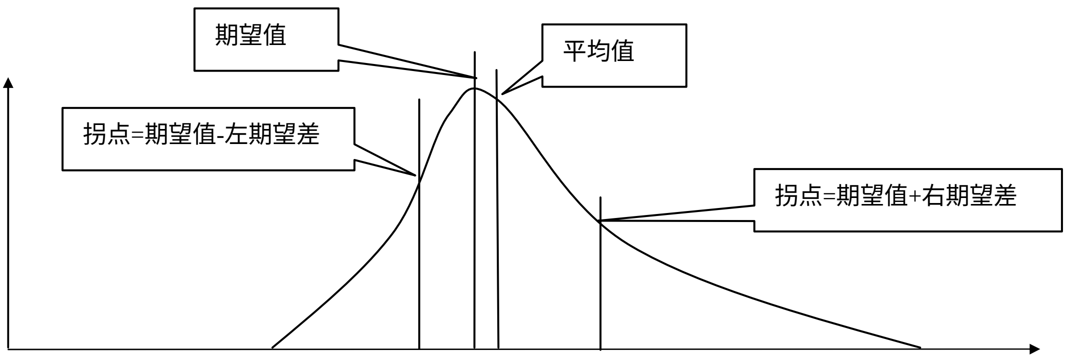

图 1 偏斜分布曲线的拐点=期望值±左右期望差示意图

2．左期望差右期望差提出的理论依据

在正态分布的情况下，期望值与平均值等价。对称性说明左期望差与右期望差相等，期望差与传统意义上的标准差完全等价。

在偏斜分布的情况下，期望值不等于平均值，呈现出不对称，这样就决定了左期望差与右期望差是不相等的。但是任意的单峰分布以期望值为中线，左右两边都可以看成是正态分布的一半，则可以应用高斯分布原理分别进行描述。推论如下：

以图 1 为例进行讨论，以期望值为中轴左边图形向右旋转 180^{O} 就成为完全意义上的正态分布。由此可以得到平均值等于期望值，也就可以从传统的计算标准差的结果中得到左期望差。

同理，以期望值为中轴右边图形向左旋转 180^{O} 就成为完全意义上的正态分布。此时的平均值同样等于期望值，也就可以从传统的计算标准差的结果中得到右期望差。

以上由不同的左右图形通过相对旋转形成两个不同的正态分布，平均值与期望值都是同一个数，但是，它们离散程度的变异指标西格玛显然不同，从图 1 中可以明显看到由左边形成的正态分布其离散程度的变异指标小于右边形成的正态分布的离散程度的变异指标。由此可以得到一个结论：在单峰的条件下任意偏斜分布以期望值为中线，两边都有属于各自不同的离散程度的变异指标。分别称为：左期望差与右期望差。

国际统计科学界的领军人物范剑青教授说：“统计的最大问题在于模型误差，局部建模的优点在于可以大大降低误差。”[4]

以上说明左期望差、右期望差的提出依据的是范剑青教授局部建模的新理念，即：任意单峰分布以期望值为中线就可以分成两个局部来进行分部描述，这样的描述所依据的是高斯分布的原理。可以从以下单峰分布的密度函数[3]得到验证：

$$f(x)=\begin{cases}\left(\dfrac{1}{\sqrt{2\pi}\sigma_-}\right)\exp\left[-\dfrac{(x-\mu)^2}{2\sigma_-^2}\right], & x\le\mu\\[2ex]\left(\dfrac{1}{\sqrt{2\pi}\sigma_+}\right)\exp\left[-\dfrac{(x-\mu)^2}{2\sigma_+^2}\right], & x\ge\mu\end{cases}$$

当左期望差 σ_- 和右期望差 σ_+ 相等时，上式就是完全意义上的高斯分布。当左期望差 σ_- 和右期望差 σ_+ 不相等时，呈偏斜分布。以上 f(x)分部函数依据的是范剑青教授局部建模的新理念。所以上式 f(x)既能描述正态分布也能描述偏斜分布。由此得出：提出左期望差 σ_- 和右期望差 σ_+ 依据的是高斯分布的原理。

3. 期望差的定义与计算方法

单峰分布密度函数的建立所引出的期望差、左期望差和右期望差的新概念，这是从统计学应用实践中客观存在的在单峰分布情况下不对称分布曲线求其拐点（见图 1）问题的背景条件下产生的。这是数学应用发展的必然结果，由此引出确定不对称分布曲线拐点的计算则是统计学自身发展的需要。所提出的期望差、左期望差和右期望差同样适用于正态分布。以下将其相应的定义完整地重述如下：

“期望差定义为：随机变量 x 与期望值 μ 的离差平方和的平均的平方根。符号记为：σ 。

左期望差定义为：小于期望值 μ 随机变量 x 与期望值 μ 的离差平方和的平均的平方根。符号记为：σ_-。

右期望差定义为：大于期望值 μ 随机变量 x 与期望值 μ 的离差平方和的平均的平方根。符号记为：σ_+。”[3]

其公式仍以图 1 为例来展开描述。

左期望差、右期望差、期望差计算方法的表达式如下：

令：随机变量频数之和为 n；

小于期望值的频数为 n_1；

大于期望值的频数为 n_2；

等于期望值的频数为 n_3。

设：期望值左边的频数为 n_-；

期望值右边的频数为 n_+；

则：$n_- = n_1 + n_3 \div 2$

$n_+ = n_2 + n_3 \div 2$

满足：$n = n_- + n_+ = n_1 + n_2 + n_3$

根据以上条件计算 σ_-、σ_+和 σ 的公式如下：

左期望差：$\sigma_- = [\sum_{i=1}^{n_-} (x_i - \mu)^2 / n_-]^{1/2}$，$x_i \leq \mu$；

右期望差：$\sigma_+ = [\sum_{i=n_-+1}^{n} (x_i - \mu)^2 / n_+]^{1/2}$，$x_i \geq \mu$；

期望差：$\sigma = [\sum_{i=1}^{n} (x_i - \mu)^2 / n]^{1/2}$。

根据以上计算公式对应计算样本期望差公式如下：

样本左期望差：$S_- = [\sum_{i=1}^{n_-} (x_i - \mu)^2 / (n_- - 0.5)]^{1/2}$，$x_i \leq \mu$；

样本右期望差：$S_+ = [\sum_{i=n_-+1}^{n} (x_i - \mu)^2 / (n_+ - 0.5)]^{1/2}$，$x_i \geq \mu$；

样本期望差：$S = [\sum_{i=1}^{n} (x_i - \mu)^2 / (n - 1)]^{1/2}$。

4. 偏斜分布两拐点间概率确定的推论

在单峰分布的条件下随机变量以期望值为中心存在左右两个期望差。期望值减左期望差是单峰分布左边曲线的拐点；期望值加右期望差是单峰分布右边曲线的拐点。当左右期望差不相等时，呈偏斜分布。如图 1 所示。

由于单峰分布包括正态分布和偏斜分布。分布曲线两拐点间的概率计算仍然需要应用高斯分布的原理来进行推论，一方面说明正态分布和偏斜分布之间的内在联系。另一方面只要确定了偏斜分布概率的计算问题则单峰分布任意区间概率的计算问题也就解决了。

推论：

仍然以图 1 为例，以期望值为中轴左边图形向右旋转 180^O 就成为左边形成的正态分布。同理，以期望值为中轴右边图形向左旋转 180^O 就成为右边形成的正态分布。此时的平均值等于期望值，也就可以从传统的计算标准差的结果中得到：

左边形成的正态分布的标准差等价左期望差。

右边形成的正态分布的标准差等价右期望差。

左边形成的正态分布的拐点为：期望值减左期望差，加左期望差是分布曲线的拐点，两拐点间的概率是 0.6826。

右边形成的正态分布的拐点为：期望值减右期望差，加右期望差是分布曲线的拐点，两拐点间的概率是 0.6826。

从上述期望差的计算公式已知：

图 1 随机变量频数之和为 n；

小于期望值的频数为 n_1；

大于期望值的频数为 n_2；

等于期望值的频数为 n_3。

期望值左边的频数为 n_-；

期望值右边的频数为 n_+；

则：$n = n_- + n_+ = n_1 + n_2 + n_3$

令：左边形成的正态分布的频数之和为 n_4；

右边形成的正态分布的频数之和为 n_5；

有：$n_4 = n_1 \times 2 + n_3$

$n_5 = n_2 \times 2 + n_3$

设：左边形成的正态分布的权数为：f_-

右边形成的正态分布的权数为：f_+；

则：$f_- = (n_1 \times 2 + n_3) \div (n_4 + n_5) = (n_1 + n_3 \div 2) \div n = n_- \div n$

$f_+ = (n_2 \times 2 + n_3) \div (n_4 + n_5) = (n_2 + n_3 \div 2) \div n = n_+ \div n$

有：$f_- + f_+ = n_- \div n + n_+ \div n = 1$

由图 1 以期望值两边形成的正态分布，期望值相同，左右期望差不同，所形成的正态分布也不同，但两个不同正态分布两拐点间的概率均为：0.6826。对应其权数，容易得图 1 两拐点间的概率为：0.6826。

即：$0.6826 f_- + 0.6826 f_+$

$= 0.6826 (n_- \div n) + 0.6826 (n_+ \div n)$

$= 0.6826 (n_- \div n + n_+ \div n)$

$= 0.6826$

还可以从另外一个角度来说明。

若随机变量 ξ 服从正态分布 $N(\mu_1, \sigma_-, \sigma_+)$ 有：$\sigma_- = \sigma_+ = \sigma$，

则 ξ 落在区间 $(\mu_1 - \sigma, \mu_1 + \sigma)$ 的概率等于 $P(|\xi - \mu_1| < \sigma) = 0.6826$；

若随机变量 ξ 服从偏斜分布 $N(\mu_2, \sigma_-, \sigma_+)$，有 $\sigma_- \neq \sigma_+ \neq \sigma$

则 ξ 落在区间 $(\mu_2 - \sigma_-, \mu_2 + \sigma_+)$ 内的概率与正态分布相同

即：区间 $(\mu_2 - \sigma_-, \mu_2 + \sigma_+)$ 的概率等于 0.6826；

以下给出证明：

已知：

偏斜分布左边的频数为 n_-；

偏斜分布右边的频数为 n_+。

有：总频数 $n = n_- + n_+$；

偏斜分布左边的比率（权数）为 n_-/n；

偏斜分布右边的比率（权数）为 n_+/n。

令：偏斜分布左边的概率为 1。

有：$P(\mu_2 - \sigma_-, \mu_2) = 0.6826$；

同理

令：偏斜分布右边的概率为 1。

有：$P(\mu_2, \mu_2 + \sigma_+) = 0.6826$；

则：$P(\mu_2 - \sigma_-, \mu_2 + \sigma_+) = P(\mu_2 - \sigma_-, \mu_2)\, n_-/n + P(\mu_2, \mu_2 + \sigma_+)\, n_+/n$

$= 0.6826\, n_-/n + 0.6826\, n_+/n$

$= 0.6826 (n_-/n + n_+/n)$

$= 0.6826$

通过以上证明得到以下推论：正态分布与偏斜分布有如下关系：

$P(\mu_1-k\sigma, \mu_1+k\sigma) = P(\mu_2-k\sigma_-, \mu_2+k\sigma_+)$ k=1，2，3，…

即：$P(\mu_1-\sigma, \mu_1+\sigma) = P(\mu_2-\sigma_-, \mu_2+\sigma_+) = 0.6826$

这里特别强调正态分布与偏斜分布的关系还有：

$$P(\mu_1-i\sigma, \mu_1+j\sigma) \neq P(\mu_2-i\sigma_-, \mu_2+j\sigma_+)$$

$i, j = 1, 2, 3, \cdots\cdots$ 且 $i \neq j$

即：$P(\mu_1-\sigma, \mu_1+2\sigma) \neq P(\mu_2-\sigma_-, \mu_2+2\sigma_+)$

这是由于偏斜分布左右两边比率（权数）不同所致。

在任意单峰分布密度函数中，当期望值等于分布中心数[5]时，且左右期望差相等时，呈现正态分布。否则为偏斜分布。

从以上正态分布区间概率推论偏斜分布区间的概率，从而确定了偏斜分布区间的概率。为其分布任意区间概率的计算提供了理论依据。

5. 期望差提出的实践意义

从以上论述过程中已经对所提出的期望差在统计实践中的意义作了一些描述，概括主要有如下几点：

1．传统意义上的标准差只能计算在正态分布条件下分布曲线的拐点。在单峰分布的条件下客观存在的偏斜分布曲线的拐点的计算问题则是由左期望差和右期望差来给以确定，从而为计算偏斜分布任意区间的概率奠定理论基础。

2．单峰分布密度函数 f(x)的数学表达式分部的左函数为单增函数其参数是左期望差；右函数为单减函数其参数是右期望差。它们是单峰分布的两个重要参数。

3．左期望差和右期望差是否相等是判定单峰分布是否对称的主要依据。

4．左期望差和右期望差的提出应用于统计分析能把随机变量数据间的差异与变动情况说得更清楚一些。

5．拓展了西格玛的内涵。[6]

6．期望差、左期望差和右期望差的提出区别了传统意义上的标准差。

6. 结论

单峰分布包括对称的正态分布和不对称的偏斜分布。

在正态分布的条件下平均值等价期望值，其分布曲线的拐点在平均值加减传统意义标准差所对应密度函数的点上。

在偏斜分布的条件下，传统意义上的标准差与偏斜分布曲线的拐点无关，在此统计背景情况下，依据高斯分布原理和范剑青教授局部建模的新理念，从单峰分布具有的特征，即：任意单峰分布以期望值为中线都可以看成是正态分布的一半，就此进行分部描述，以提出左期望差、右期望差的新概念，从而确定偏斜分布曲线的拐点，为计算偏斜分布任意区间的概率奠定理论基础。由此可推广应用到统计分析和统计推断的实践中去。

参考文献

[1] 孔建新 孔建勇《建立单峰分布密度函数的探讨》[OL]. 中国科技论文在线， 2008.04.24.
[2] 孔建新 孔建勇《偏斜分布密度函数的提出与推论》[OL]. 中国科技论文在线， 2008.03.05.
[3] 孔建新 孔璐 何光伟《有关单峰分布应用规范问题的研究》[OL]. 中国科技论文在线， 2008.11.3
[4] lixing《范剑青 把数学作为解决社会问题的工具》[OL] 经济学论坛-中国经济学教育科研网 2008-2-13

[5] 孔建新 孔璐 何光伟《关于对单峰分布中心数位置特征的初步讨论》[OL]. 中国科技论文在 2008.10.15
[6] 孔建新 李飞荣 孔粤《关于对西格玛内涵拓展的研究》[OL]. 中国科技论文在线， 2008.12.10

The expectation deviation bring forward of statistical background and significance

Konglu[1] Kongyue[2] Kongjianxin[3]
1Zhongshan Torch Development Zone in Guangdong Province Vocational and Technical College .Zhongshan City, 528437 E-mail：yudianlulu@163.com
2Yunnan Nationalities University School of Continuing Education
Kunming City in Yunnan Province, 650031 E-mail：baobao_723@yahoo.com.cn
3 Yunnan Hui Zhe Zhe Hai, 2-24-6 statistics scientific group 654211
Email: kongfanjx@163.com

Abstract: At unimodal distribution of conditions under expectations deviation is the random variable and expectations numerical of discrete degree of with the Variability indicator. As the unimodal distribution include the symmetric of normal distribution and the asymmetric of partial distribution. However the traditional sense of the standard deviation is the random variable and average numerical discrete degree of the variability indicator. A skewed distribution when the distribution curve to determine the inflection point associated with the expectations deviation of Related.With the traditional sense of standard deviation no relationship.
This is the expectation deviation bring forward of statistical background. The meaning is: to determine the distribution of any single peak curve inflection point for the purpose of calculating any period of its laid a theoretical basis for probability.

Key words: Unimodal distribution；Skewed distribution；Expectations deviation；Left、Right expectations deviation；Inflection Point