单峰分布应用实例的演算

孔璐[1] 何光伟[2] 孔建新[3]

1 暨南大学工业工程管理专业在职研究生 广州市（510632）

E-mail：yudianlulu@163.com

2 广东药学院中山校区 中山市（528458）

E-mail：weiweihonest@163.com

3 云南驰宏公司会泽技术监督处质量统计组 云南会泽者海镇（654211）

E-mail：kongfanjx@163.com

摘 要：本文以社会经济的统计对象：居民收入统计资料的实例，以说明呈现的频数服从单峰分布。应用单峰分布密度函数的数学模型来描述其分布曲线，能够达到拟合的优度，从而提高统计推断的精确性。通过演算其分布任意区间的概率来验证单峰分布在实践中的意义。

关键词：单峰分布；居民收入；数学模型；实例；演算

中图分类号：O211.66

1. 引言

社会经济统计对象有关数量指标的频数客观存在多种分布形态。在一般情况下普遍的情况是：在单峰的条件下客观存在对称正态分布和不对称偏斜分布两种。在此统计的背景下，单峰分布密度函数数学模型[1]的提出，统一了这两种分布形态。解决了以往只用正态分布来拟合描述不对称的偏斜分布所以带来的统计误差。

单峰分布统一的两种分布形态所要达到的目的是：建立一个既能描述正态分布又能描述偏斜分布的密度函数。以满足这一类不同形态分布的情况下能达到精确描述以降低统计误差的要求。

本文以某地区居民收入频数分布的统计资料为例，应用单峰分布密度函数的数学模型来描述其分布曲线，通过计算此分布任意给定的一个区间的概率来验证单峰分布在实践中的意义。

2. 居民收入分布状况

在社会经济统计对象有关数量指标的频数服从单峰分布是一普遍现象。如：人口年龄分布，男或女的身高和体重的分布，某地区高考成绩分布，居民收入分布以及各类产品质量指标的频数分布等，基本服从单峰分布，可能呈现对称正态分布或者不对称偏斜分布。本文以居民收入统计资料为例进行演算。

根据 2007 年某地区随机抽查访问（自由职业人员）和调查（各类企业员工收入）的统计资料（为本课题研究的需要进行的访问调查，虽无权威性但有一定的代表性，仅供参考）以推测模拟某地区居民收入数据分布状况列表如下（见表 1 第(1)，(2)，(3)，(7)列）：

表 1 某地区居民收入模拟数据分布计算统计表

组距	组中值	众数左边（频数单位：千人）				众数右边（频数单位：千人）			
元	元	频数	离差	离差平方	加权离差平方	频数	离差	离差平方	加权离差平方
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
≤500	450	4	-1100	121	484	—	—	—	—
(500～600]	550	6	-1000	100	600	—	—	—	—
(600～700]	650	14	-900	81	1134	—	—	—	—
(700～800]	750	16	-800	64	1024	—	—	—	—
(800～900]	850	21	-700	49	1029	—	—	—	—
(900～1000]	950	25	-600	36	900	—	—	—	—
(1000～1100]	1050	28	-500	25	700	—	—	—	—
(1100～1200]	1150	30	-400	16	480	—	—	—	—
(1200～1300]	1250	36	-300	9	324	—	—	—	—
(1300～1400]	1350	41	-200	4	164	—	—	—	—
(1400～1500]	1450	50	-100	1	50	—	—	—	—
(1500～1600]	1550	58	0	0	0	58	0	0	0
(1600～1700]	1650	—	—	—	—	52	100	1	52
(1700～1800]	1750	—	—	—	—	49	200	4	196
(1800～1900]	1850	—	—	—	—	41	300	9	369
(1900～2000]	1950	—	—	—	—	39	400	16	624
(2000～2100]	2050	—	—	—	—	33	500	25	825
(2100～2200]	2150	—	—	—	—	30	600	36	1080
(2200～2300]	2250	—	—	—	—	29	700	49	1421
(2300～2400]	2350	—	—	—	—	28	800	64	1792
(2400～2500]	2450	—	—	—	—	25	900	91	2025
(2500～2600]	2550	—	—	—	—	21	1000	100	2100
(2600～2700]	2650	—	—	—	—	17	1100	121	2057
(2700～2800]	2750	—	—	—	—	13	1200	144	1872
(2800～2900]	2850	—	—	—	—	11	1300	169	1859
(2900～3000]	2950	—	—	—	—	9	1400	196	1764
(3000～3100]	3050	—	—	—	—	9	1500	225	2025
(3100～3200]	3150	—	—	—	—	8	1600	256	2048
(3200～3300]	3250	—	—	—	—	5	1700	289	1445
(3300～3400]	3350	—	—	—	—	5	1800	324	1620
(3400～3500]	3450	—	—	—	—	4	1900	391	1444
(3500～3600]	3550	—	—	—	—	3	2000	400	1200
(3600～3700]	3650	—	—	—	—	2	2100	441	882
(3700～3800]	3750	—	—	—	—	2	2200	484	968
(3800～3900]	2850	—	—	—	—	1	2300	169	169
(3900～4000]	3950	—	—	—	—	1	2400	576	576
＞4000	4050	—	—	—	—	1	2500	625	625

续表 1 某地区居民收入模拟数据分布计算统计表

组距	组中值	众数左边（频数单位：千人）				众数右边（频数单位：千人）			
元	元	频数	离差	离差平方	加权离差平方	频数	离差	离差平方	加权离差平方
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
合计	—	329	—	—	6889	496	—	—	31038
众数 σ^2	—	—	—	—	22.9633	—	—	—	66.4625
众数 σ	—	—	—	—	479 元	—	—	—	815 元

注：表中的频数栏单位为千人，离差平方栏和加权离差平方栏单位为万元。

根据统计资料计算得居民收入的平均值为 1787 元。其分布频数最高的峰值（众数）为 1550 元。

左西格玛 $\sigma_- = 479$

右西格玛 $\sigma_+ = 815$

通过统计计算得出，平均数不等于众数，左西格玛 σ_-与右西格玛 σ_+不相等。根据单峰分布的性质 1，频数呈现的是不对称的偏斜分布。

见图 1。

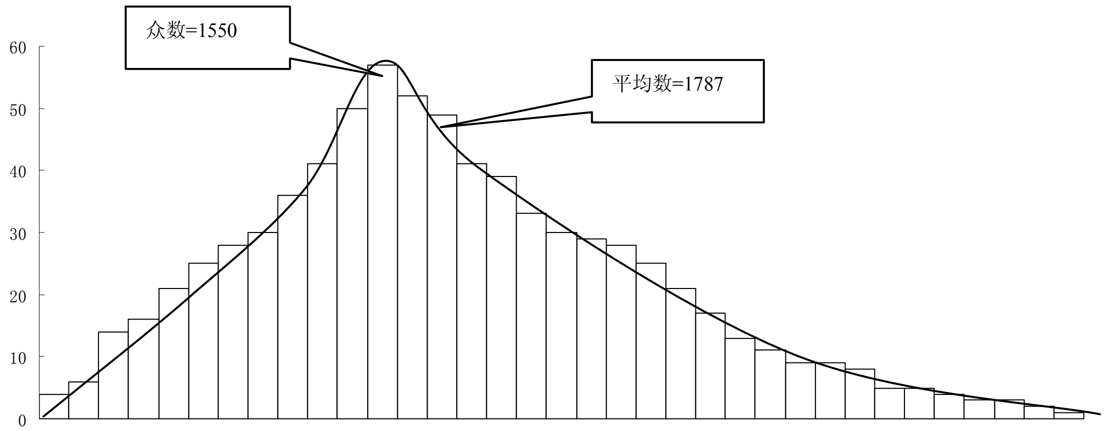

图 1 居民收入分布直方图与分布曲线示意图

3. 居民收入分布曲线呈现偏斜分布的数学表达

从居民收入统计资料的频数统计直方图展示得出：居民收入频数服从偏斜分布[2]。根据单峰分布密度函数公式：

$$f(x)=\begin{cases}\left(\dfrac{1}{\sqrt{2\pi}\sigma_-}\right)\exp\left[-\dfrac{(x-M_0)^2}{2\sigma_-^2}\right], & x\leqslant M_0\\[2ex]\left(\dfrac{1}{\sqrt{2\pi}\sigma_+}\right)\exp\left[-\dfrac{(x-M_0)^2}{2\sigma_+^2}\right], & x\geqslant M_0\end{cases}$$ [1]

从以上统计资料已知：$\mu\neq M_0$、$\sigma_-\neq\sigma_+$、$n_-\neq n_+$。根据单峰分布的性质 1，居民收入频数服从偏斜分布。

设：小于众数的频数之和为：n_1；

大于众数的频数之和为：n_2；

等于众数的频数为：n_3。

令：众数与众数左边的频数为：n_-；

众数与众数右边的频数为：n_+。

则：$n_- = n_1 + n_3 \div 2$；

$n_+ = n_2 + n_3 \div 2$。

满足：$n = n_1 + n_2 + n_3 = n_- + n_+$ 。

根据表 1 资料，有（频数单位：千人）：

$n_1 = 271$， $n_2 = 438$， $n_3 = 58$

$n_- = n_1 + n_3 \div 2 = 271 + 58 \div 2 = 300$；

$n_+ = n_2 + n_3 \div 2 = 438 + 58 \div 2 = 467$。

通过对模拟居民收入分布统计表的计算结果得到：

众数（单峰分布数学期望值）：$M_0 = 1550$，

众数左方差：$\sigma_-^2 = \frac{1}{n_-}\sum_{i=1}^{n_-}(x_i - M_0)^2 = \frac{1}{300}\sum_{i=1}^{n_-}(x_i - 1550)^2 = 229633$

众数右方差：$\sigma_+^2 = \frac{1}{n_+}\sum_{i=n_-+1}^{n}(x_i - M_0)^2 = \frac{1}{467}\sum_{i=n_-+1}^{n}(x_i - 1550)^2 = 664625$

众数左西格玛： $\sigma_- = [\frac{1}{n_-}\sum_{i=1}^{n_-}(x_i - M_0)^2]^{1/2} = 229633^{1/2} = 479$

众数右西格玛：$\sigma_+ = [\frac{1}{n_+}\sum_{i=n_-+1}^{n}(x_i - M_0)^2]^{1/2} = 664625^{1/2} = 815$

居民模拟数据资料的分布曲线（见图 1），其分布曲线得到如下分布函数：

$$f(x) = \begin{cases} (\frac{1}{\sqrt{2\pi}\,479})\exp[-\frac{(x-1550)^2}{2\times 229633}], & x \leqslant 1550 \\ (\frac{1}{\sqrt{2\pi}\,815})\exp[-\frac{(x-1550)^2}{2\times 664625}], & x \geqslant 1550 \end{cases}$$

由于整个分布区间的概率等于 1，有：

偏斜分布左函数的概率：$P_-\{x \leqslant M_0\} = \int_{-\infty}^{M_0} f_-(t)\,dt = n_-/n$ ；

偏斜分布右函数的概率：$P_+\{x \geqslant M_0\} = \int_{M_0}^{+\infty} f_+(t)\,dt = n_+/n$ 。

满足： $P\{-\infty < X < +\infty\} = \int_{-\infty}^{M_0} f_-(t)\,dt + \int_{M_0}^{+\infty} f_+(t)\,dt = n_-/n + n_+/n = 1$

为了以后方便计算偏斜分布任意区间概率，将偏斜分布左函数的概率 n_-/n，偏斜分布右函数的概率 n_+/n，分别作为各自两边的权数，分别记为 f_1，f_2。

根据单峰分布概率密度函数的数学模型及表 1 统计资料，现将此分布任意给定区间进行概率的计算过程演算如下：

4. 偏斜分布给定区间概率计算方法的演算

根据以上公式及以上计算出的已知结果，分布任意区间的概率计算过程如下：

已知：

期望值：$M_0 = 1550$

偏斜分布左函数的频数：$n_- = 300$

偏斜分布右函数的频数：$n_+ = 467$

左西格玛：$\sigma_- = 479$

右西格玛：$\sigma_+ = 815$

分部左函数频数的比率（权数）：$f_1 = n_- \div n = 300 \div 767 = 0.39$

分部右函数频数的比率（权数）：$f_2 = n_+ \div n = 467 \div 767 = 0.61$

根据表 1 居民收入模拟数据服从参数为 $M_0 = 1550$，$\sigma_- = 479$，$\sigma_+ = 815$ 的偏斜分布。设 $\xi \sim N(M_0, \sigma_-, \sigma_+)$ 即，$\xi \sim N(1550, 479, 815)$。试求居民收入在区间[1000，2000]的概率。已知 $M_0 = 1550$，以众数为中轴将所求概率的区间[1000，2000]分为 [1000，1550]，[1550，2000] 两个区间来分别计算。（注：这两个区间来分别是正态分布的一半）

所求概率分左右两部分的分布函数表示为：

$P_-\{1000 \leqslant \xi \leqslant 1550\} = F(1550) - F(1000)$

$P_+\{1550 \leqslant \xi \leqslant 2000\} = F(2000) - F(1550)$

设所求区间各点值分别为：$x_1 = 1000$，$x_2 = 1550$，$x_3 = 2000$

分布左函数，设 $\xi \sim N(M_0, \sigma_-^2)$，所求概率 $P_-\{1000 \leqslant \xi \leqslant 1550\}$。

取变换 $\eta = \dfrac{\xi - M_0}{\sigma_-}$，

设：$P_{正-}$ 为按正态分布计算分布左函数的概率，

$$P_{正-}\{x_1 \leqslant \xi \leqslant x_2\} = P_{正-}\left\{\frac{\chi_1 - M_0}{\sigma_-} < \frac{\xi - M_0}{\sigma_-} < \frac{\chi_2 - M_0}{\sigma_-}\right\}$$

$$= \left\{\frac{\chi_1 - M_0}{\sigma_-} < \eta < \frac{\chi_2 - M_0}{\sigma_-}\right\}$$

$$= \Phi\left(\frac{\chi_2 - M_0}{\sigma_-}\right) - \Phi\left(\frac{\chi_1 - M_0}{\sigma_-}\right)$$

将已知数代入，得：

$$P_{正-}\{1000 \leqslant \xi \leqslant 1550\} = \Phi\left(\frac{1550 - 1550}{479}\right) - \Phi\left(\frac{1000 - 1550}{479}\right)$$

$$= \Phi(0) - \Phi(-1.1482)$$

$$= \Phi(0) - [1 - \Phi(1.1482)]$$

$$= 0.5 - [1 - 0.8746]$$

$$= 0.5 - 0.1254$$

$$= 0.3746$$

分布右函数，设 $\xi \sim N(M_0, \sigma_+^2)$，所求概率 $P_+\{1550 \leqslant \xi \leqslant 2000\}$。

取变换 $\eta = \dfrac{\xi - M_0}{\sigma_+}$，

设：$P_{正+}$ 为按正态分布计算分布右函数的概率，

$$P_{正+}\{x_2 \leqslant \xi \leqslant x_3\} = P_{正+}\left\{\frac{\chi_2 - M_0}{\sigma_+} < \frac{\xi - M_0}{\sigma_+} < \frac{\chi_3 - M_0}{\sigma_+}\right\}$$

$$= \left\{\frac{\chi_2 - M_0}{\sigma_+} < \eta < \frac{\chi_3 - M_0}{\sigma_+}\right\}$$

$$= \Phi\left(\frac{\chi_3 - M_0}{\sigma_+}\right) - \Phi\left(\frac{\chi_2 - M_0}{\sigma_+}\right)$$

将已知数代入，得：

$$P_{正+}\{1550 \leqslant \xi \leqslant 2000\} = \Phi\left(\frac{2000-1550}{815}\right) - \Phi\left(\frac{1550-1550}{815}\right)$$

$$= \Phi(0.5521) - \Phi(0)$$

$$= 0.7096 - 0.5$$

$$= 0.2096$$

以上计算偏斜分布各自左右两部分的概率，仅仅占各自部分的一半，还需各自乘以 2。以下给出理由：

前面已经提出，在偏斜分的条件下，众数减众数左标准差，众数加众数右标准差，是此分布曲线的拐点。以众数为中线左曲线的拐点至众数区间的概率为 0.3413，由于把左函数看作是正态分布的一半，所以左曲线拐点至众数区间的概率为等于 0.3413×2=0.6826，它是占左边整个区间的概率。同理，右也一样。

令：左边函数概率为 1。

设：P_-为偏斜分布左函数概率。

则：$P_-\{1000 \leqslant \xi \leqslant 1550\} = P_{正-}\{1000 \leqslant \xi \leqslant 1550\} \times 2 = 0.3746 \times 2 = 0.7492$

即：0.7492 是 $P_-\{1000 \leqslant \xi \leqslant 1550\}$占左函数区间的概率。

令：右边函数概率为 1。

设：P_+为偏斜分布右函数概率。

则：$P_+\{1550 \leqslant \xi \leqslant 2000\} = P_{正+}\{1550 \leqslant \xi \leqslant 2000\} \times 2 = 0.2096 \times 2 = 0.4192$

即：0.4192 是 $P_+\{1550 \leqslant \xi \leqslant 2000\}$占左函数区间的概率。

根据已知偏斜分布左右函数频数的比率（权数）：

分部左函数频数的比率（权数）：$f_1 = n_- \div n = 300 \div 767 = 0.3911$

分部右函数频数的比率（权数）：$f_2 = n_+ \div n = 467 \div 767 = 0.6099$

按比率调整计算偏斜分布的左右函数的实际概率，方法及计算过程如下：

设：$P_{偏-}$为按比率调整计算偏斜分布左函数的实际概率，

$$P_{偏-}\{1000 \leqslant \xi \leqslant 1550\} = P_-\{1000 \leqslant \xi \leqslant 1550\} \times f_1$$

$$= 0.7492 \times 0.3911$$

$$= 0.2930$$

设：$P_{偏+}$为按比率调整计算偏斜分布右函数的实际概率，

$$P_{偏+}\{1550 \leqslant \xi \leqslant 2000\} = P_+\{1550 \leqslant \xi \leqslant 2000\} \times f_2$$

$$= 0.2096 \times 0.6099$$

$$= 0.1278$$

设：$P_{偏}$为按比率调整计算偏斜分布整体函数的实际概率，

$$P_{偏}\{1000 \leqslant \xi \leqslant 2000\} = P_{偏-}\{1000 \leqslant \xi \leqslant 1550\} + P_{偏+}\{1550 \leqslant \xi \leqslant 2000\}$$

$$= 0.2930 + 0.1278$$

$$= 0.4208$$

求得某地区居民收入在区间[1000，2000]中的概率为 0.4208。

通过以上实例计算过程可以简概为：在计算偏斜分布任意区间概率的过程中，把偏斜分布看作是两个对应的不同的正态分布，它们的平均数都等于偏斜分布的期望值众数。用计算正态分布概率的方法对应不同的权数就可以求得偏斜分布任意区间的概率。

通过以上实际计算还验证了单峰分布的性质[1]，当众数等于平均数，且众数左标准差等于右标准差时，便还原为正态分布任意区间概率的计算。从以上偏斜分布任意区间概率的计

算还说明高斯分布的原理始终贯穿于其中。而且运用偏斜分布概率密度函数的公式也完全适用于计算正态分布任意区间的概率，此时权数 $f_1 = f_2$ 。

5. 结论

通过以上实例的计算过程说明，若仅用正态分布的数学模型来描述不对称的偏斜分布显然存在一定的统计误差，因为在偏斜分布的条件下平均数±σ不是分布曲线的拐点，所以在平均数±σ区间的概率不等于 0.6826。但是在计算偏斜分布任意区间的概率时，则需要把偏斜分布转化为不同的两个正态分布，用计算正态分布概率的方法来计算偏斜分布的概率。

从实例的计算验证了单峰分布数学模型的建立不但可以精确描述偏斜分布，而且还可以精确描述正态分布。

以上对某地区模拟的居民收入频数的分布状况用单峰分布的数学模型来描述并进行任意区间概率的计算，其结果表明：单峰分布曲线密度函数的建立完全能解决描述居民收入等一类经济与社会现象呈现偏斜分布状况。如：产品质量指标的频数分布，人口年龄分布，高考成绩分布等等，都能以此数学模型来进行描述和进行统计推断。

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］http：//www.Paper.edu.cn 2008. 4. 28

[2] 孔建新 孔建勇《偏斜分布密度函数的提出与推论》［OL］http：//www.Paper.edu.cn 2008. 3. 5

Unimodal distribution application example calculation

Konglu[1] Heguangwei[2] Kongjianxin[3]

1 Jinan University industrial engineering management professionals working graduate students
Guangdong Province Guangzhou City, 528404
Email: yudianlulu@163.com

2 Guangdong College of Pharmacy, Zhongshan District, Zhongshan City 528458
E-mail：weiweihonest@163.com

3 Chi-Hung companies Yunnan Huize Statistics Section of the technical supervision
Zhe Hai, Hui Zhe Yun Nan 654211, P.R. China,
Email: kongfanjx@163.com

Abstract： This article by social economy statistical object: The resident income statistics material's example, showed presents frequency obedience unimodal distribution. Applies the unimodal distribution density function the mathematical model to describe its distribution curve, can achieve the fitting the goodness of fit, thus enhancement statistical inference accuracy. Through calculates its distribution random sector probability to confirm unimodal distribution significance in reality.

Key word： Unimodal distribution; Resident income; Mathematical model; Example; Calculation