·监测·

中国MSM人群HIV感染者病毒载量抽样调查数据分布特征及数据转换研究

斗智　陈军　江震　宋炜路　徐杰　吴尊友

102206 北京，中国疾病预防控制中心性病艾滋病预防控制中心预防干预室（斗智、陈军、江震、宋炜路、徐杰）；102206 北京，中国疾病预防控制中心性病艾滋病预防控制中心（吴尊友）

通信作者：江震，Email：jiangzhen812@126.com

DOI：10.3760/cma.j.issn.0254-6450.2017.11.011

【摘要】 目的　了解MSM人群的HIV感染者（MSM感染者）人群病毒载量（PVL）数据分布特征，拟合分布函数，探讨评价PVL的合适参数。**方法**　病毒载量（VL）检测限设定为≤50拷贝/ml。描述PVL的一般分布特征，结合Box-Cox转换和正态性检验，根据PVL数据转换后的分布特征，拟合稳定分布函数，并进行拟合优度检验。**结果**　PVL原始数据为偏态分布，变异系数（CV）为622.24%，经Box-Cox转换，转换参数（λ）最优值=-0.11，为多峰分布；VL原始值>检测限的PVL经Box-Cox数据转换，λ最优值=0，为对数转换，偏态厚尾特征，不满足正态分布，拟合稳定分布函数（α=1.70，β=-1.00，γ=0.78，δ=4.03），呈稳定分布。**结论**　PVL原始值存在截尾、非正态分布的特征，变异度较大；当VL原始值≤检测限的截尾数据占总体比例较大时，不宜用检测限的1/2代替；VL原始值>检测限的PVL对数值为稳定分布，适合用*M*和*IQR*来描述集中趋势和离散趋势。

【关键词】　艾滋病病毒；病毒载量；分布特征

基金项目：国家科技重大专项（2012ZX10001007005）

Data distribution and transformation in population based sampling survey of viral load in HIV positive men who have sex with men in China　*Dou Zhi, Chen Jun, Jiang Zhen, Song Weilu, Xu Jie, Wu Zunyou*

Division of Prevention and Intervention, National Center for AIDS/STD Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China (Dou Z, Chen J, Jiang Z, Song WL, Xu J); National Center for AIDS/STD Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China (Wu ZY)

Corresponding author: Jiang Zhen, Email: jiangzhen812@126.com

【Abstract】 Objective　To understand the distribution of population viral load (PVL) data in HIV infected men who have sex with men (MSM), fit distribution function and explore the appropriate estimating parameter of PVL. **Methods**　The detection limit of viral load (VL) was ≤ 50 copies/ml. Box-Cox transformation and normal distribution tests were used to describe the general distribution characteristics of the original and transformed data of PVL, then the stable distribution function was fitted with test of goodness of fit. **Results**　The original PVL data fitted a skewed distribution with the variation coefficient of 622.24%, and had a multimodal distribution after Box-Cox transformation with optimal parameter (λ) of -0.11. The distribution of PVL data over the detection limit was skewed and heavy tailed when transformed by Box-Cox with optimal λ=0. By fitting the distribution function of the transformed data over the detection limit, it matched the stable distribution (SD) function (α=1.70, β=-1.00, γ=0.78, δ=4.03). **Conclusions**　The original PVL data had some censored data below the detection limit, and the data over the detection limit had abnormal distribution with large degree of variation. When proportion of the censored data was large, it was inappropriate to use half-value of detection limit to replace the censored ones. The log-transformed data over the detection limit fitted the SD. The median (*M*) and inter-quartile ranger (*IQR*) of log-transformed data can be used to describe the centralized tendency and dispersion tendency of the data over the detection limit.

【Key words】 Human immunodeficiency virus; Viral load; Distribution characteristics
Fund program: National Major Science and Technology Project of China (2012ZX10001007005)

HIV病毒载量(viral load, VL)是艾滋病防治工作中一项重要的实验室检测指标,在艾滋病临床诊断和抗病毒治疗以及艾滋病科研工作中得到广泛应用[1]。HIV感染者人群病毒载量(population viral load, PVL)数据分布特征对统计分析存在影响,VL分布一般呈连续偏态分布[2]。此外,HIV感染者在接受抗病毒治疗后,其VL值≤检测限(或者无法检测),在临床上或实验室被称为检测限以下数据,这类截尾数据增加了统计分析的难度。了解PVL的分布特征,应用合理的统计方法,是分析PVL影响因素的重要前提。有研究指出,PVL原始分布以及对数转换均呈偏态分布[2-3]。本研究分析了我国2013年16个大城市MSM人群HIV感染者(MSM感染者)VL数据,为了准确描述PVL分布特征,探讨描述PVL的合理参数,为艾滋病防治及科研工作提供参考依据。

资料与方法

1. 资料来源:国家科技重大专项子项目"MSM人群扩大检测扩大治疗降低新发感染试点项目"目的在于有效治疗MSM感染者、预防继发传播、降低HIV新发感染。其中,以MSM感染者VL水平作为评价干预效果的指标之一。数据来源于北京、上海、南京、杭州、武汉、重庆、昆明、西安、广州、深圳、南宁、乌鲁木齐、哈尔滨、长春、成都和天津共16个大城市MSM感染者VL抽样调查。调查时间为2013年5月至2015年12月。在基线、干预后1年、干预后2年的3个时间点随机抽取4 050例MSM感染者,收集其VL检测数据,比较MSM感染者VL的变化趋势。本研究选择2013年基线数据来分析PVL分布特征。

2. 研究方法:

(1)VL检测限的处理:各城市的VL检测试剂及设备并不完全一致,存在各自不同的VL检测限,本研究以罗氏诊断产品有限公司COBAS Taqman VL试剂的检测限为参照,校正其他试剂,以50拷贝/ml作为VL检测限参考值。如果VL原始值≤检测限,则取其检测限的1/2作为其VL近似值[4]。

(2)描述和分析PVL分布特征:由于VL数据呈偏态分布特征,运用Box-Cox方法正态转换并做正态性检验,对于VL值>检测限的VL数据拟合稳定分布(stable distribution, SD)函数,并进行拟合优度检验。

数据正态转换方法较多,如倒数转换、对数转换、平方根转换、平方根反正弦转换等[5]。当数据满足对数正态分布,资料的标准差与均值之比接近时,可以借助对数转换,使数据满足正态分布,见公式(1)。当数据满足泊松分布,或者方差与均数正相关时,可以进行平方根转换,使数据满足正态分布,见公式(2)。当数据为样本率,呈二项分布时,可以通过平方根反正弦转换,见公式(3)。当数据两端变异较大时,可以通过倒数转换,减小极端值的影响,见公式(4)。

$$Y=\log(y) \quad (1)$$

$$Y=\sqrt{(y)} \quad (2)$$

$$Y=\frac{1}{\sin}\sqrt{(y)} \quad (3)$$

$$Y=1/y \quad (4)$$

上述转换可以被统一起来,形成一个转换组,即Box-Cox转换[6],其表达式:

$$Y=\begin{cases}\dfrac{y^{\lambda}-1}{\lambda} & \lambda\neq 0\\ \log(y) & \lambda=0\end{cases} \quad (5)$$

$$L=-\frac{v}{2}\ln(S_y^2)+(\lambda-1)\frac{v}{n}\sum\ln(y) \quad (6)$$

式中,S_y^2表示转换后新变量的方差,n为样本量,v为自由度,当L为最大值时的转换参数(λ)值,为最优值。

由公式(5)可知,当$\lambda=0$时等于做对数转换,$\lambda=-1$时等于做倒数转换,$\lambda=0.5$时等于做平方根转换。对λ的估计,有两种方法,最大似然法和Bayes法。本研究采取最大似然法估计转换参数(λ)最优值,见公式(6)。使用SAS 9.3软件对数据进行Box-Cox转换,设定λ变化范围(-3,3),步长值0.01,软件可以根据最大似然法依次计算,得出λ最优值。

SD是具有偏态厚尾特征的单峰概率分布,可以描述现实中许多随机效应[7-9]。SD有4个参数(α,β,γ,δ),其中$\alpha\in[0,2]$,$\beta\in[-1,1]$,$\gamma\in(-\infty,\infty)$,$\delta\in(0,\infty)$。$\alpha$为特征参数,表示分布曲线尾部的厚度,$\alpha$越小,尾部越厚,$\alpha$越接近2表示越接近正态分布。$\beta$表示偏度,$\beta=0$表示分布对称,$\beta<0$表示分布呈右偏,$\beta>0$表示分布呈左偏。$\gamma$为位置参数,$\delta$是刻度参数。SD的参数估计一般有3种方法:分位数法、样本函数法和最大似然函数法,精度越来越高。SD函数表达式:

$$\phi(t)=\begin{cases}\exp\left\{-\gamma^{\alpha}|t|^{\alpha}\left[1+i\beta\left(\tan\frac{\pi\alpha}{2}\right)(\gamma|t|)^{1-\alpha}-1\right]+i\delta t\right\} & (\alpha\neq 1)\\ \exp\left\{-\gamma|t|\left[1+i\beta\frac{\pi}{2}(\sin t)\right]\right\} & \\ \ln\gamma+i\delta t\Big\} & (\alpha=1)\end{cases} \quad (7)$$

本研究采用Stable软件[10]进行拟合SD函数，估计4个参数（α,β,γ,δ）。对SD的拟合优度检验，是结合2种方法综合判断[7]，一是用3种不同的参数估计方法，估计（α,β,γ,δ），直观比较3种估计方法得到的参数值之间的差异是否较大，如果不大，说明拟合较好；另一种是利用Stable软件拟合的理论数据，在Matlab软件中作实际值和理论值的累计概率分布图，直观判断两条曲线是否较好吻合，若基本吻合，说明拟合结果理想。

结　果

2013年基线调查样本量4 050例MSM感染者，实际共收集2 879例，完成率为71.1%。VL原始值＞检测限的有1 593例（55.3%，1 593/2 879），VL原始值≤检测限的有1 286例（44.7%，1 286/2 879）。

1. PVL原始值分布特征：

（1）PVL原始值总体分布特征（表1）：PVL原始值总体的变异系数（coefficient of variation, CV）为622.24%，偏度系数为22.96，正态性检验Kolmogorov-Smirnov（K-S）值＝0.44，P＜0.05，频数分布直方图显示数据呈偏态，变异性较大。见图1。

（2）VL原始值＞检测限的PVL原始值分布特征（表1）：由于VL原始值＜检测限的占比为44.7%，如果取1/2值可能会造成数据偏态分布，只取VL原始值＞检测限的数据分析。VL原始值＞检测限的PVL原始值的CV为474.37%，偏度系数为19.71，K-S值为0.42，P＜0.05，频数分布直方图显示数据呈偏态，变异性较大。见图2。

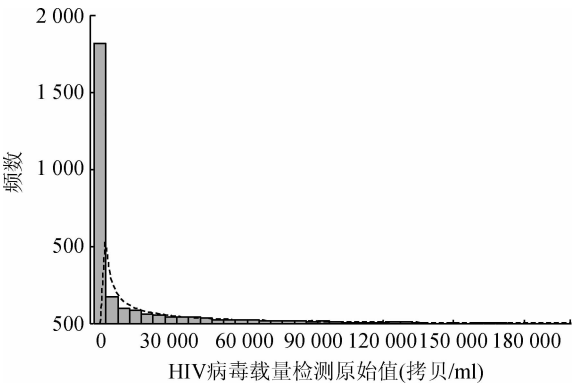

图1　人群病毒载量原始值总体分布直方图及拟合曲线

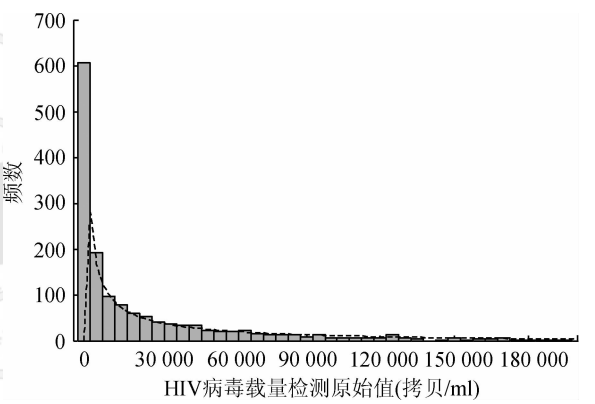

图2　病毒载量原始值＞检测限的人群病毒载量原始值分布直方图和拟合曲线

2. PVL原始值Box-Cox转换后的分布特征：

（1）PVL原始值的Box-Cox转换结果及分布特征（表1）：经Box-Cox转换，得到λ最优值＝–0.11。CV＝36.96%，偏度系数为–0.22，正态性检验K-S值0.21，P＜0.05，频数直方图显示数据呈偏态和多峰

表1　中国MSM感染者病毒载量抽样调查PVL原始值分布特征

统计学指标	PVL原始值分布特征			
	总体特征	经Box-Cox转换后分布特征	VL原始值＞检测限的PVL原始值分布特征	VL原始值＞检测限的PVL原始值经Box-Cox转换后分布特征
样本量(例)	2 879	2 879	1 593	1 593
$\bar{x}$值	35 350.00	0.56	55 514.00	3.64
s	219 959.00	0.21	263 341.00	1.23
M值	54	0.64	7 290	3.86
最小值	5	0.17	16	1.21
最大值	7 770 000	0.84	7 770 000	6.89
变异系数(%)	622.24	36.96	474.37	33.69
偏度系数	22.96	–0.22	19.71	–0.39
峰度系数	674.83	–1.65	508.50	–0.82
K-S值[a]	0.44	0.21	0.42	0.09

注：[a]表示P＜0.05；PVL指人群病毒载量；VL指病毒载量

分布，变异性较大，见图3。

（2）VL原始值＞检测限的PVL原始值Box-Cox转换后结果及分布特征（表1）：经Box-Cox转换，得到λ最优值＝0，为对数转换。CV＝33.69%，偏度系数为−0.39，正态性检验K-S值0.09，P＜0.05，频数分布直方图显示分布呈现偏态和厚尾的特征，不满足正态分布，见图4。

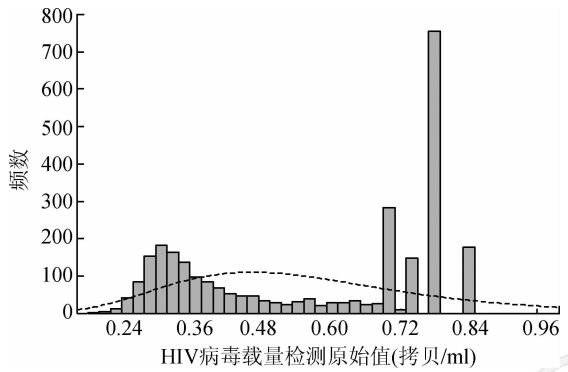

图3 人群病毒载量原始值Box-Cox转换后分布直方图和拟合曲线

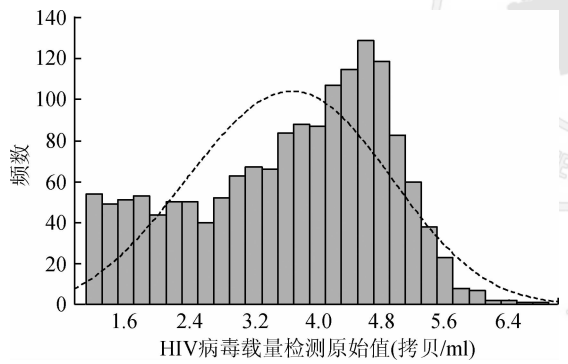

图4 病毒载量原始值大于检测限的人群病毒载量原始值Box-Cox转换后分布直方图和拟合曲线

3. VL原始值＞检测限的PVL对数转换后SD分布函数（表2）：PVL原始值转换后呈现多峰分布，无法进行参数拟合。单独对VL原始值＞检测限的PVL数据进行对数转换后，数据呈现单峰、偏态、厚尾的特征，可能满足SD。拟合SD，分别用分位数法、样本函数法和极大似然法进行参数估计，其中极大似然法可以估计参数的95%*CI*。由表2可知，3种方法估计的参数差异较小。α＝1.7说明分布呈现厚尾特征，β＝−1说明分布为右偏，符合图4的分布特征。以最大似然函数法估计的参数值来拟合函数，对实际值和理论值作累计概率分布图检验拟合情况，左侧尾部分离较大，其他部分理论值和实际值基本吻合，VL原始值＞检测限的PVL值对数转换后，PVL对数值满足SD分布。见图5。

表2 人群病毒载量原始值＞检测限的病毒载量对数转换后SD分布函数

参数估计方法	分位数法	样本函数法	极大似然法(95%*CI*)
α	1.81	1.70	1.70(1.67～1.73)
β	−1.00	−1.00	−1.00(−0.98～1.02)
γ	0.93	0.89	0.78(0.76～0.80)
δ	4.08	4.08	4.03(4.01～4.05)

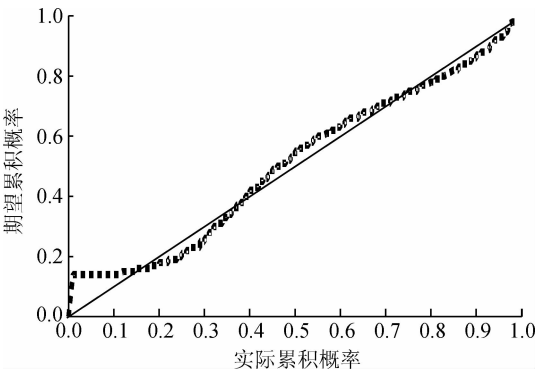

图5 病毒载量原始值大于检测限的人群病毒载量对数转换后SD分布累计概率

讨　论

1. PVL原始值分布特征及转换效果：本研究发现，我国MSM的PVL原始值为连续性偏态分布，变异性较大，表现为左端截尾，与国外研究结论基本一致[3,11]。VL原始值＞检测限的PVL原始值CV小于总体原始值的CV。Box-Cox正态转换后，VL原始值＞检测限的PVL对数值表现为单峰、偏态、厚尾分布特征，PVL原始值转换后，表现为多峰分布特征，主要是由检测限的1/2值代替截尾值造成。为避免截尾值造成数据信息丢失，一些研究用检测限的1/2值代替截尾值[4]，当截尾值占总体的比例较小时，这种方法会有一定效果。当截尾值占总体的比例较大时[12]，会造成PVL总体数据的变异性增大，影响数据转换效果。

本研究中，VL原始值＞检测限的PVL对数值，满足SD，国外的一些研究指出这种为偏正态分布[13]（skew-normal distribution，SND），表现为单峰、偏态、厚尾的特征。这种分布在实际情况中较为常见，如市场上的股票收益率[14-15]。在本研究中，可能是因为影响VL的众多因素，并不是随机和相互抵消的，有一些影响因素强度较大[9]，如治疗等一些人为干扰因素，从而造成了偏态分布。

2. PVL的参数选择：PVL数据可以分为VL原始值≤检测限和VL原始值＞检测限两个部分，分别描述PVL分布特征。对VL原始值≤检测限的数据

做1/2处理后，总体PVL原始值变异度较大，转换后呈现多峰分布，不适宜用一个指标来概述PVL的分布特征。在PVL水平表述中，建议列出VL原始值≤检测限的截尾数据占总体的比例。VL原始值>检测限数据，对数转换后呈稳定分布，可取对数值*M*和*IQR*描述集中趋势和离散趋势[16]。

3. PVL的统计分析：采取两种统计方法分析PVL数据，一是处理为分类变量，采用二分类非条件logistic回归[2]，按照病毒抑制成功的标准（≤1 000拷贝/ml），所有VL原始值≤检测限的情况，认为其VL抑制成功；二是作为连续性变量，采用校正的Tobit模型分析。VL原始值>检测限的PVL进行对数转换，VL原始值≤检测限的PVL处理为截尾值，这样数据整体表现为左端截尾，偏态的分布特征。传统的Tobit模型可以处理截尾数据，其参数估计方法为最大似然法，这种方法的前提是潜变量的正态性和方差齐性，为克服PVL数据的偏态性，有研究指出采用含有半参数混合效应的Tobit模型来分析[10]。

4. 研究局限性和意义：本研究验证了国内MSM人群VL数据的分布特征与国际研究近似，为今后分析VL数据选择合理的统计方法提供了依据。局限性在于样本人群仅包括MSM感染者，并存在VL值缺失数据。

总之，PVL原始值存在截尾、非正态分布的特征，变异度较大；当VL原始值≤检测限的截尾数据占总体比例较大时，不宜用检测限1/2值代替，要报告截尾数据占总体的比例；VL原始值>检测限的PVL对数值为SD，适合用其对数值的*M*和*IQR*描述其集中趋势和离散趋势。

志谢　本文得到国家科技重大专项MSM人群艾滋病干预研究课题组的16个现场工作组成员（卢红艳、曾吉、王娟、于茂河、徐鹏、郭伟、梅淑娟、李雪静、李一、闫红梅、刘岩琳、庄鸣华、宁镇、沈晓沛、还锡萍、闫红静、张敏、朱正平、潘晓红、王懿、罗艳、张兴亮、蒋洪林、汤恒、刘普林、李艳、徐慧芳、程伟彬、钟斐、刘少础、蓝光华、陈怡、农全兴、李恬、龚毅、何勤英、范双凤、吴国辉、欧阳琳、闵向东、章任重、梁军、常文辉、贾华、卫晓丽、吴明旭、倪明建、李凡、李瑞兰、王新迪和王云霞等）的支持，以及徐晓玉、任仙龙、陈军、曹巍和Nanci Nanyi Zhang的大力协助

利益冲突　无

参 考 文 献

[1] 蒋岩，潘品良，李敬云，等. HIV-1病毒载量检测及质量保证指南（2007版）[EB/OL]. 北京：中国疾病预防控制中心性病艾滋病预防控制中心.（2008-08-02）[2016-06-30]. http://www.chinaaids.cn/jszn/200808/t20080802_1099125.htm.
Jiang Y, Pan PL, Li JY, et al. The guideline of HIV-1 viral load testing and quality assurance (2007) [EB/OL]. Beijing: National Center for AIDS/STD Control and Prevention, China CDC. (2008-08-02) [2016-06-30]. http://www.chinaaids.cn/jszn/200808/t20080802_1099125.htm.

[2] Grando LJ, Machado DC, Spitzer S, et al. Viral coinfection in the oral cavity of HIV-infected children: relation among HIV viral load, CD_4^+ T lymphocyte count and detection of EBV, CMV and HSV [J]. Braz Oral Res, 2005, 19(3): 228-234. DOI: 10.1590/S1806-83242005000300013.

[3] Dagne GA. Bayesian inference for skew-normal mixture models with left-censoring [J]. J Biopharm Stat, 2013, 23(5): 1023-1041. DOI: 10.1080/10543406.2013.813517.

[4] CDC. Guidance on community viral load: a family of measures, definitions, and method for calculation [DB/OL]. (2011-08-31) [2016-06-30]. http://stacks.cdc.gov/view/cdc/28147.

[5] 王慧文，潘秀丹. 变量变换的通用公式[J]. 中国卫生统计，1993，10(3)：43-44.
Wang HW, Pan XD. A general formula of the variable transformation [J]. Chin J Health Stat, 1993, 10(3): 43-44.

[6] Sakia RM. The Box-Cox transformation technique: a review [J]. J R Stat Soc, 1992, 41(2): 169-178. DOI: 10.2307/2348250.

[7] Bergström H. On some expansions of stable distribution functions [J]. Ark Mat, 1952, 2(4): 375-378. DOI: 10.1007/BF02591503.

[8] Kida S. Log-stable distribution and intermittency of turbulence [J]. J Phys Soc Jpn, 1991, 60(1): 5-8. DOI: 10.1143/JPSJ.60.5.

[9] 顾娟，茆诗松. 稳定分布的参数估计[J]. 应用概率统计，2002，18(4)：342-346. DOI: 10.3969/j.issn.1001-4268.2002.04.002.
Gu J, Mao SS. The estimation of the parameter of stable distribution [J]. Chin J Appl Probab Stat, 2002, 18(4): 342-346. DOI: 10.3969/j.issn.1001-4268.2002.04.002.

[10] Nolan JP. Stable distributions: models for heavy tailed data [DB/OL]. (2009-05-13) [2016-06-30]. http://fs2.american.edu/jpnolan/www/stable/stable.html.

[11] Dagne GA, Huang YX. Bayesian semiparametric mixture Tobit models with left censoring, skewness, and covariate measurement errors [J]. Stat Med, 2013, 32(22): 3881-3898. DOI: 10.1002/sim.5799.

[12] Hornung RW, Reed LD. Estimation of average concentration in the presence of nondetectable values [J]. Appl Occup Environ Hyg, 1990, 5(1): 46-51. DOI: 10.1080/1047322X.1990.10389587.

[13] Bandyopadhyay D, Lachos VH, Castro LM, et al. Skew-normal/independent linear mixed models for censored responses with applications to HIV viral loads [J]. Biom J, 2012, 54(3): 405-425. DOI: 10.1002/bimj.201000173.

[14] 卢方元. 中国股市收益率分布特征研究[J]. 中国管理科学，2004，12(6)：18-22. DOI: 10.3321/j.issn.1003-207X.2004.06.004.
Lu FY. A research on distribution characteristics of stock market returns in China [J]. Chin J Manage Sci, 2004, 12(6): 18-22. DOI: 10.3321/j.issn.1003-207X.2004.06.004.

[15] 王建华，王玉玲，柯开明. 中国股票收益率的稳定分布拟合与检验[J]. 武汉理工大学学报，2003，25(10)：99-102. DOI: 10.3321/j.issn.1671-4431.2003.10.029.
Wang JH, Wang YL, Ke KM. The stable distributions fitting and testing in stock returns of China [J]. J Wuhan Univ Technol, 2003, 25(10): 99-102. DOI: 10.3321/j.issn.1671-4431.2003.10.029.

[16] 方积乾. 卫生统计学[M]. 北京：人民卫生出版社，2010.
Fang JQ. Biostatistics [M]. Beijing: The People's Medical Publishing House, 2010.

（收稿日期：2017-01-20）
（本文编辑：王岚）