Advances in Applied Mathematics 应用数学进展, 2022, 11(8), 5509-5517
Published Online August 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.118580

Hans汉斯

基于拟合优度检验的无症状感染者统计模型分析

——以上海疫情为例

祖煜然*，赵建昕

海军潜艇学院，山东 青岛

收稿日期：2022年7月8日；录用日期：2022年8月2日；发布日期：2022年8月12日

摘 要

新冠病毒的不断变异和大规模人群接种疫苗等原因，使得疫情爆发地区中的无症状感染者增多。分析和掌握无症状感染者的增长规律，对制定防疫政策和精准防疫措施有重要的参考意义。本文对2022年上海疫情中无症状感染者数据进行统计分布拟合，分析拟合分布的偏态、峰值和分位点，发现伽马分布和对数正态分布与真实分布较为近似，并通过EDF统计量对拟合分布作拟合优度检验，结果显示伽马分布检验*p*值较大，检验通过率较高，其拟合效果较好。

关键词

无症状感染者，统计分布，EDF统计量，拟合优度检验

Statistical Model Analysis of Asymptomatic Patients Based on Goodness of Fit Test

—Taking the Shanghai Epidemic as an Example

Yuran Zu*, Jianxin Zhao

Navy Submarine Academy, Qingdao Shandong

Received: Jul. 8th, 2022; accepted: Aug. 2nd, 2022; published: Aug. 12th, 2022

Abstract

Due to the continuous mutation of novel coronavirus and mass vaccination of the population, the

*通讯作者。

文章引用: 祖煜然, 赵建昕. 基于拟合优度检验的无症状感染者统计模型分析[J]. 应用数学进展, 2022, 11(8): 5509-5517.
DOI: 10.12677/aam.2022.118580

number of asymptomatic patients in regional outbreaks has increased. Analyzing and mastering the growth pattern of asymptomatic infections has important reference significance for formulating epidemic prevention policies and precise epidemic prevention measures. This paper performs statistical distribution fitting on the data of asymptomatic patients in the Shanghai epidemic in 2022, analyzes skewness, peaks, and quantiles of fitted distributions, it is found that the gamma distribution and lognormal distribution are more approximate to the true distribution. In this paper, EDF statistic is used to test the goodness of fit of the fitting distributions. The results show that the p-value of the gamma distribution test is large, and the test pass rate is high, indicating that the fitting effect is good.

Keywords

Asymptomatic Patients, Statistical Distributions, EDF Statistics, Goodness of Fit Test

1. 引言

2022 年新型冠状病毒(2019-nCoV)在多地区爆发，但随着病毒的不断变异、疫苗接种率的提升以及疫情管控措施的加强，相对于 2020 年疫情发展，无症状感染者成为主要病毒感染人群。现阶段新冠疫情中，无症状感染者也具有较强的传染性，同时具有隐匿性，使得无症状感染者的人数增长速度要快于确诊病例，对于疫情防控而言是一个新的挑战。因此分析无症状感染者增长的统计分布，研究其传播规律，对当前疫情走势判断、防控政策的制定有着重要的意义。

统计分布，是用来描述随机现象的基本工具，是数理统计学的基础，任何统计方法都离不开统计分布的概念和各种具体分布的性质[1]。根据确定的统计分布，可以研究事件发展变化情况，对未知事件的发生进行预测。统计分布分包括离散型统计分布和连续型统计分布，本文将利用连续型分布研究无症状感染者的增长情况。常见的连续型统计分布有均匀分布、正态分布、对数正态分布、伽马分布和威布尔分布等，本文通过参数估计的方法拟合出无症状感染者数据的多种连续型统计分布，并对每一种分布进行拟合优度检验，找到最优的统计模型。拟合优度检验是用来检验观测数与依照某种假设或分布模型计算得到的理论数之间一致性的一种统计假设检验，以便判断该假设或模型是否与实际观测数相吻合，是评价用已知分布拟合现实数据优劣的一种方法，有着广泛的实际应用[2]。拟合优度检验包括作图法与回归方法、χ^2 型检验和 EDF 型检验等，借助 EDF 统计量检验拟合出的分布是本文的主要研究方法。

2. 统计分布拟合

2.1. 分布模型

2.1.1. 正态分布模型

若随机变量 X 的概率密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

则称随机变量 X 服从均值为 μ，方差为 σ^2 的正态分布，记作 $X\sim N\left(\mu,\sigma^2\right)$，其中

$$\hat{\mu}=\bar{X}\text{，}\quad \hat{\sigma}=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 \tag{1}$$

分别为 μ 和 σ^2 的一个无偏估计，那么可以用分布 $N\left(\hat{\mu},\hat{\sigma}^2\right)$ 近似随机变量 X 的真实分布。

2.1.2. 对数正态分布模型

若随机变量 X 的取值为正数，且 $\ln X\sim N\left(\mu,\sigma^2\right)$，则称随机变量 X 服从对数正态分布，其概率密度函数为

$$f\left(x\right)=\begin{cases}\dfrac{1}{x\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{\left(\ln x-\mu\right)^2}{2\sigma^2}}, & x>0\\ 0, & x\le 0\end{cases}$$

在对数正态分布的实际应用中，一般取 μ 和 σ^2 的极大似然估计作为 μ 和 σ^2 估计量，分别为

$$\hat{\mu}_{MLE}=\frac{1}{n}\sum_{i=1}^{n}\ln X_i\text{，}\quad \hat{\sigma}_{MLE}=\frac{1}{n}\sum_{i=1}^{n}\left(\ln X_i-\frac{1}{n}\sum_{i=1}^{n}\ln X_i\right)^2 \tag{2}$$

2.1.3. 伽马分布模型

若随机变量 X 的概率密度函数为

$$f\left(x\right)=\begin{cases}\dfrac{\beta^{\alpha}}{\Gamma\left(\alpha\right)}x^{\alpha-1}\mathrm{e}^{-\beta x}, & x>0\\ 0, & x\le 0\end{cases}$$

其中 $\Gamma\left(\alpha\right)=\int_0^{+\infty}x^{\alpha-1}\mathrm{e}^{-x}\mathrm{d}x$，$\alpha$ 为形状参数，β 为尺度参数，则称随机变量 X 服从伽马分布，记作 $X\sim G\left(\alpha,\beta\right)$，伽马分布的期望和方差分别为

$$E\left(X\right)=\frac{\alpha}{\beta}\text{，}\quad Var\left(X\right)=\frac{\alpha}{\beta^2} \tag{3}$$

用样本均值 $\frac{1}{n}\sum_{i=1}^{n}X_i$ 代替真实均值，样本方差 $\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$ 代替真实方差，根据公式(3)可以计算出参数 α 和 β 的估计量

$$\hat{\alpha}=\frac{\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2}{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}\text{，}\quad \hat{\beta}=\frac{\frac{1}{n}\sum_{i=1}^{n}X_i}{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2} \tag{4}$$

2.1.4. 威布尔分布模型

若随机变量 X 的概率密度函数为

$$f\left(x\right)=\begin{cases}\dfrac{k}{\lambda}\left(\dfrac{x}{\lambda}\right)^{k-1}\mathrm{e}^{-\left(\frac{x}{\lambda}\right)^k}, & x>0\\ 0, & x\le 0\end{cases}$$

其中 $\lambda>0$ 为尺度参数，$k>0$ 为形状参数，则称随机变量 X 服从威布尔分布，记作 $X\sim W\left(\lambda,k\right)$。用实际数据进行分布拟合时，一般使用的参数估计为

$$\hat{k}=\frac{1.2}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\ln X_i-\frac{1}{n}\sum_{i=1}^{n}\ln X_i\right)}},\quad \hat{\lambda}=\exp\left(\frac{1}{n}\sum_{i=1}^{n}\ln X_i+\frac{0.572}{\hat{k}}\right) \tag{5}$$

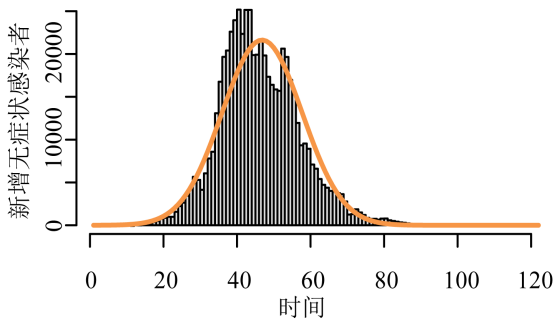
2.2. 数据拟合

本文收集了上海 2022 年 3 月 1 日至 6 月 30 日共 122 天的无症状感染者日增长数据，数据来源于上海市卫生健康委官方网站[3]。为了研究上海新冠疫情无症状感染者的增长规律，本节对日增长数据进行统计分布拟合。在疫情蔓延过程中，日增长数据会出现偶然的波动，为了消除这种偶然因素，提高模型的拟合优度，本节还采用移动平均的方法拟合数据，其中以 7 天作为平均长度。在数据拟合过程中，将无症状感染者出现的日期作为随机变量，为了便于描述，把 3 月 1 日记作上海疫情第 1 天，6 月 30 日记作上海疫情第 122 天，则随机变量 X 的取值为 $1,2,3,\cdots,122$。由公式(1)、(2)、(4)、(5)，根据无症状感染者日增长数据和 7 天移动平均数据，计算出四个统计分布的参数估计，具体如表 1。

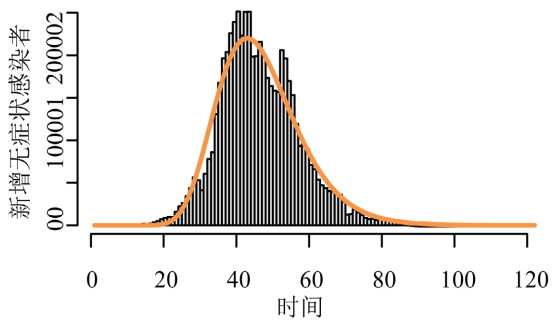
Table 1. Probability distribution parameter estimates for two classes of data
表 1. 两类数据的统计分布参数估计值

分布	参数	日平均估计值	7 天移动平均估计值	分布	参数	日平均估计值	7 天移动平均估计值
$X\sim N(\mu,\sigma^2)$	$\hat{\mu}$	46.84	43.84	$X\sim G(\alpha,\beta)$	$\hat{\alpha}$	17.97	15.07
	$\hat{\sigma}^2$	119.10	122.88		$\hat{\beta}$	0.38	0.34
$\ln X\sim N(\mu,\sigma^2)$	$\hat{\mu}$	3.82	3.75	$X\sim W(\lambda,k)$	$\hat{k}$	4.45	4.14
	$\hat{\sigma}^2$	0.06	0.07		$\hat{\lambda}$	51.14	48.10

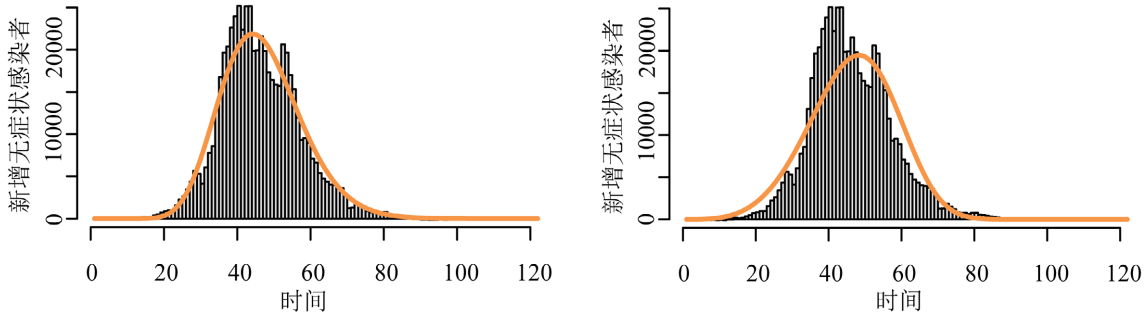
为了便于观察无症状感染者增长数，本文的分布图采用的是频数分布图。图 1 和图 2 的直方图分别展示了无症状感染者日增长数和 7 天移动平均增长数，图中曲线表示四种拟合分布的频数变化。由图可以发现，无症状感染者前期增长速度较快，呈正偏态分布。在拟合的分布曲线中，对数正态分布和伽马分布是正偏态的，而威布尔分布是负偏态的；同时，对数正态分布和伽马分布对尾部数据拟合较好，能较好地表示疫情后期无症状感染者的增长情况；从峰值上看，四个分布的峰值都低于真实峰值，但对数正态分布和伽马分布的峰值与真实数据的峰值最为接近；从峰值出现的时间上看，真实数据峰值出现的时间与对数正态分布和伽马分布峰值出现的时间近似，而正态分布和威布尔分布的峰值出现的时间要晚。通过以上分析可知，对数正态分布和伽马分布对两类数据拟合的效果较好，而正态分布和威布尔分布拟合效果较差。



(a) 正态分布拟合

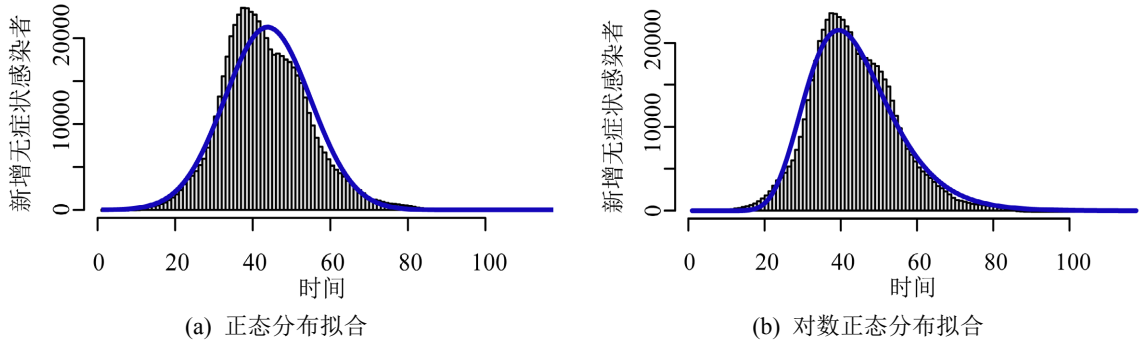


(b) 对数正态分布拟合

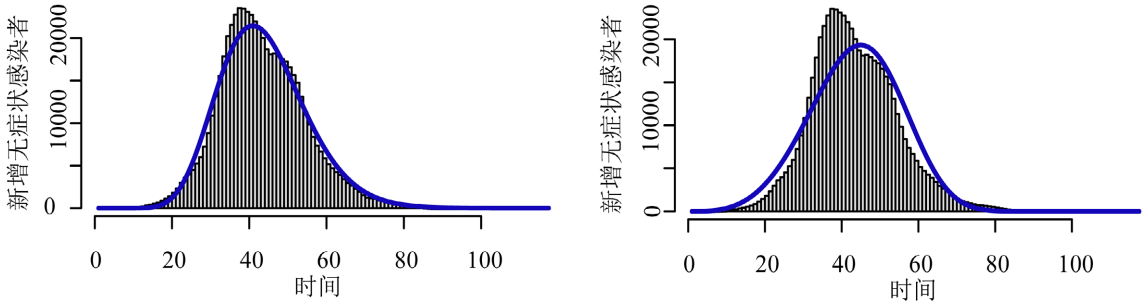


(c) 伽马分布拟合　　(d) 威布尔分布拟合

Figure 1. Frequency plots of fitted distributions for daily growth data
图 1. 日增长数据拟合分布频数图



(a) 正态分布拟合　　(b) 对数正态分布拟合



(c) 伽马分布拟合　　(d) 威布尔分布拟合

Figure 2. Frequency plots of fitted distributions for 7-day moving average growth data
图 2. 7 天移动平均增长数据拟合分布频数图

中国精算师协会传染病数学模型研究小组在研究 2003 年非典疫情北京数据数学模型时，利用分位点评价分布拟合效果[4]，本文运用了同样的方法对分布进行拟合评价。表 2 和表 3 中的数据是两类数据真实分布和拟合分布的分位点，通过表格数据可以分析疫情发展情况与时间的关系，例如，真实数据的 50%分位点为 39，表示第 1 天至 39 天无症状感染者增长总数占 122 天总感染人数的 50%，此时疫情发展处于中期。从表中数据可以发现，四个拟合分布的 25%、50%、75%分位点与真实分布的分位点都很接近，伽马分布的这三个分位点与真实分布分位点完全相同。若 99.9%分位点作为此次疫情即将结束的时间点，则真实分布中第 85 天进入疫情结束阶段，正态分布和威布尔分布将此时间点提前 7 至 8 天，对数正态分布推迟 8 至 9 天，而伽马分布仅推迟 1 至 2 天。在 95%、99%分位点上，同样是伽马分布表现得最好。综上所述，对于疫情发展情况，四个分布在疫情前中期都能够较好地判断，而在疫情发展后期，伽马分布判断较为准确。

Table 2. Quantile of the fitted distributions for daily growth data
表 2. 日增长数据拟合分布的分位点

分布	25% 分位点	50% 分位点	75% 分位点	95% 分位点	99% 分位点	99.9% 分位点
真实分布	39	46	54	66	77	87
正态分布	39	47	54	65	72	80
对数正态分布	39	46	54	68	80	96
伽马分布	39	46	54	66	76	88
威布尔分布	39	47	55	65	72	79

Table 3. Quantile of the fitted distributions for 7-day moving average growth data
表 3. 7 天移动平均增长数据拟合分布的分位点

分布	25% 分位点	50% 分位点	75% 分位点	95% 分位点	99% 分位点	99.9% 分位点
真实分布	36	43	51	63	74	85
正态分布	36	44	51	62	70	78
对数正态分布	35	42	51	66	79	97
伽马分布	36	43	51	64	74	87
威布尔分布	36	44	52	63	70	77

3. EDF 拟合优度检验

EDF 型拟合优度检验是建立在经验分布函数基础上的检验方法，常见的检验方法有 KS 检验、CvM 检验和 AD 检验。设 $X_1, X_2, \cdots, X_n$ 是一组来源于连续分布 F 的独立样本，EDF 型检验统计量检验的问题为

$$H_0 : F = F_0; H_1 : F \neq F_0$$

其中， F_0 为确定的理论分布函数。

3.1. 检验方法

3.1.1. KS 检验

Kolmogorov 和 Smirnov 提出了 KS 检验[5] [6]，KS 检验统计量为

$$D_n = \sqrt{n} \sup_{-\infty<x<\infty} \left| F_n(x) - F_0(x) \right|,$$

$$D_n^+ = \sqrt{n} \sup_{-\infty<x<\infty} \left(F_n(x) - F_0(x) \right), D_n^- = \sqrt{n} \sup_{-\infty<x<\infty} \left(F_0(x) - F_n(x) \right)$$

其中 F_n 为经验分布函数。当样本给定时，将样本按照从小到大排序得到 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$，其统计量可以表示为

$$D_n = \sqrt{n} \max_{1 \le i \le n} \left[\left(\frac{i}{n} - F_0\left(X_{(i)}\right) \right), \left(F_0\left(X_{(i)}\right) - \frac{i-1}{n} \right) \right] \tag{6}$$

3.1.2. CvM 检验

CvM 型检验统计量[7]的一般形式为

$$\omega_n^2 = n\int_{-\infty}^{\infty}\left[F_n(x)-F_0(x)\right]^2 g(x)\mathrm{d}F_0(x) \tag{7}$$

当取 $g(x)=1$ 时，便得到了 CvM 检验统计量

$$W_n^2 = n\int_{-\infty}^{\infty}\left[F_n(x)-F_0(x)\right]^2 \mathrm{d}F_0(x) \tag{8}$$

当样本给定时，CvM 检验统计量可以表示为

$$W_n^2 = \sum_{i=1}^{n}\left[F_0\left(X_{(i)}\right)-\frac{2i-1}{2n}\right]^2+\frac{1}{12n} \tag{8}$$

3.1.3. AD 检验

Anderson 和 Darling 提出了 AD 检验统计量[8]，当(7)式中 $g(x)=\dfrac{1}{x(x-1)}$ 时，可以得到 AD 检验统计量

$$A_n^2 = n\frac{\int_{-\infty}^{\infty}\left[F_n(x)-F_0(x)\right]^2}{F_0(x)\left[1-F_0(x)\right]}\mathrm{d}F_0(x)$$

当样本给定时，AD 检验统计量可以表示为

$$A_n^2 = -n-\frac{1}{n}\sum_{i=1}^{n}\left[(2i-1)\ln\left[F_0\left(X_{(i)}\right)\right]+(2n+1-2i)\ln\left[1-F_0\left(X_{(i)}\right)\right]\right] \tag{9}$$

3.2. 检验 p 值的计算方法

在进行拟合优度检验时，本文利用 p 值来确定是否拒绝原假设。对于样本 $X_1,X_2,\cdots,X_n$ 和拟合分布 F_0，EDF 型检验统计量检验 p 值的 Monte-Carlo 模拟计算步骤如下：

1) 将样本按升序排列得到 $X_{(1)},X_{(2)},\cdots,X_{(n)}$，再带入公式(6)、(8)、(9)中，分别计算出 KS 统计量 D^0，CvM 统计量 W^0，AD 统计量 A^0。

2) 运用 Monte-Carl 模拟方法，生成 n 个服从 $U[0,1]$ 分布的随机变量 $U_1,U_2,\cdots,U_n$。

3) 随机变量 $U_1,U_2,\cdots,U_n$ 升序排列得到 $U_{(1)},U_{(2)},\cdots,U_{(n)}$，令 $F_0\left(X_{(i)}\right)=U_{(i)}$，带入公式(6)、(8)、(9)中，计算三种 EDF 统计量 D,C,W。

4) 重复步骤 2)和 3) M 次(一般 M 大于 1000)，得到 $D_1,D_2,\cdots,D_n;W_1,W_2,\cdots,W_n;A_1,A_2,\cdots,A_n$。则 KS、CvM 和 AD 统计量的检验 p 值分别为

$$p_{KS}=\frac{\#\left(D>D^0\right)}{M},p_{CvM}=\frac{\#\left(W>W^0\right)}{M},p_{AD}=\frac{\#\left(A>A^0\right)}{M}$$

3.3. 检验分析

本文收集的样本量超过 59 万，这些样本不能严格地服从拟合出的四个分布，若将所有的样本进行拟合优度检验，那么四个分布都不能通过检验，而本文所研究的问题是选择一个近似的分布代替真实分布，所以需要减少检验的样本量。本文采用的方法是将所有的样本分组平均，每组的样本量为 N，那么总共分为 $H=[N/T]$ 组，其中 T 为所有样本的数量，每组样本取均值生成新的样本 $Y_1,Y_2,\cdots,Y_H$，则 Y_k $(1\le k\le H)$ 表示 122 天中前 $N(k-1)+1$ 到 Nk 个被感染者出现时间的期望。

表 4 和表 5 中显示的是两类数据拟合分布的 EDF 拟合优度检验的 p 值，在 p 值计算过程中取 $M=10000$。两表中数据表明，当 N 越小时，检验 p 值越小，这是因为用作检验的样本点越多，拟合优度检验越苛刻，

要求分布拟合的优度越高，则拟合分布越不容易通过检验。而无论 N 取何值，伽马分布的三种检验 p 值是最大的，而威布尔分布是最小的。取检验水平 $\alpha = 0.05$，当 $N = 500$ 时，从表 4 中数据可以发现，对数正态分布通过 CvM 检验，伽马分布通过了 CvM 和 AD 检验，而另外两个分布没有通过任何检验，表 5 中只有伽马分布通过了所有检验。通过以上分析可知，伽马分布的拟合效果最好。

Table 4. EDF tests p-values of fitted distributions for daily growth data
表 4. 日增长数据拟合分布的 EDF 检验 p 值

N	检验	正态分布	对数正态分布	伽马分布	威布尔分布	N	检验	正态分布	对数正态分布	伽马分布	威布尔分布
	KS	0.001	0.009	0.033	0.000		KS	0.239	0.513	0.707	0.099
500	CvM	0.008	0.070	0.188	0.000	2000	CvM	0.272	0.603	0.813	0.071
	AD	0.006	0.030	0.130	0.000		AD	0.271	0.547	0.823	0.037
	KS	0.023	0.138	0.246	0.006		KS	0.799	0.967	0.988	0.626
1000	CvM	0.069	0.268	0.484	0.006	5000	CvM	0.757	0.947	0.992	0.321
	AD	0.054	0.189	0.445	0.002		AD	0.850	0.936	0.997	0.303
	KS	0.102	0.375	0.539	0.041		KS	0.981	0.999	1.000	0.918
1500	CvM	0.195	0.456	0.668	0.027	10,000	CvM	0.962	0.999	1.000	0.656
	AD	0.195	0.361	0.669	0.012		AD	0.985	0.997	1.000	0.679

Table 5. EDF tests p-values of fitted distributions for 7-day moving average growth data
表 5. 7 天移动平均增长数据拟合分布的 EDF 检验 p 值

N	检验	正态分布	对数正态分布	伽马分布	威布尔分布	N	检验	正态分布	对数正态分布	伽马分布	威布尔分布
	KS	0.001	0.006	0.062	0.000		KS	0.326	0.484	0.795	0.213
500	CvM	0.018	0.036	0.241	0.000	2000	CvM	0.334	0.460	0.861	0.114
	AD	0.011	0.012	0.166	0.000		AD	0.352	0.408	0.871	0.074
	KS	0.050	0.107	0.337	0.023		KS	0.864	0.938	0.996	0.758
1000	CvM	0.125	0.175	0.548	0.013	5000	CvM	0.840	0.881	0.998	0.422
	AD	0.110	0.101	0.491	0.005		AD	0.906	0.855	0.997	0.419
	KS	0.181	0.280	0.595	0.116		KS	0.991	0.998	1.000	0.969
1500	CvM	0.260	0.335	0.750	0.054	10,000	CvM	0.981	0.990	1.000	0.765
	AD	0.269	0.255	0.746	0.028		AD	0.994	0.989	1.000	0.781

4. 总结

对拟合分布图进行分析可知，伽马分布、对数正态分布与真实分布的偏态方向一致，从而使得这两个分布在峰值位置上与真实分布差距不大，同时伽马分布在尾部数据拟合和分位点数值的接近程度上表

现尤为突出。在所有的 EDF 拟合优度检验中，伽马分布的检验 p 值最大，说明样本来源于伽马分布最易被接受，由此判断该分布的拟合效果最好。通过以上不同角度的分析，发现对于上海 2022 年 3 月至 6 月的疫情中无症状感染者，其增长模型适合用伽马分布进行拟合，若某地发生类似疫情，可以利用伽马分布分析预测疫情的发展情况，从而制定合理的防疫政策。

参考文献

[1] 方开泰, 许建伦. 统计分布[M]. 北京: 科学出版社, 1987.

[2] 杨振海, 程维虎, 张军舰. 拟合优度检验[M]. 北京: 科学出版社, 2011.

[3] 上海市卫生健康委. 新闻发布[EB/OL]. https://wsjkw.sh.gov.cn/xwfb/index.html, 2022-07-01.

[4] 商业新知. 传染病数学模型研究——基于 2003 年非典疫情北京数据[EB/OL]. https://www.shangyexinzhi.com/article/511180.html, 2020-02-18.

[5] Kolmogorov, A. (1933) Sulla Determinazione Empirica di Una Legge di Distribuzione. *Giornale dell'Instituto Italiano degli Attuari*, **4**, 83-91.

[6] Smirnoff, N. (1939) Sur les écarts de la Courbe de Distribution Empirique. *Matematicheskii Sbornik*, **48**, 3-26.

[7] Cramér, H. (1928) On the Composition of Elementary Errors. *Scandinavian Actuarial Journal*, **1928**, 141-180. https://doi.org/10.1080/03461238.1928.10416872

[8] Anderson, T.W. and Darling, D.A. (1952) Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, **23**, 193-212. https://doi.org/10.1214/aoms/1177729437