Distribution Pattern Analysis of Daily Confirmed COVID-19 Cases Across Selected Countries and Indian Epidemic Waves

Anita Sahu* and Jagdish Prasad

Department of Statistics, School of Applied Sciences, Amity University, Jaipur, Rajasthan 303002, India

Abstract: Daily confirmed COVID-19 cases have shown highly heterogeneous epidemic trajectories across different countries. The timing of the outbreak, the intensity of spread, the peak size, the duration of waves, and the reporting pattern varied substantially from one country to another. Due to this heterogeneity, the statistical analysis of daily confirmed cases is important not only to describe the epidemic burden but also to understand how cases are distributed over time.

In this study, the distribution pattern of COVID-19 daily cases is analysed for six countries: India, the United States, Brazil, the United Kingdom, South Africa, and Japan. The daily confirmed COVID-19 case data are taken from the Our World in Data COVID-19 dataset. The study period is restricted to a common analytical window from 30 January 2020 to 20 February 2022. After preprocessing, 753 daily observations are obtained for each country. For cross-country comparison, daily cases per million population are used, while India-specific wave analysis is performed using raw daily confirmed case counts.

The analysis evaluates descriptive distributional behaviour and parametric distribution fitting using selected candidate distributions, including Gamma, Generalised Gamma, Weibull, Log-Normal, Log-Logistic, Inverse Gaussian, Exponential, and Normal distributions. The results show that daily COVID-19 case distributions are right-skewed in all selected countries. The mean value is greater than the median value in all countries, showing that high-incidence days strongly affect the average case count. Country-wise distribution fitting shows that no single distribution is best for all countries. Log-Logistic distribution is selected for India, the United States and Japan; Inverse Gaussian distribution for Brazil; Generalised Gamma distribution for South Africa; and Log-Normal distribution for the United Kingdom.

For India, the wave-wise analysis shows that the three epidemic waves have different distributional behaviour. Wave 1 is best approximated by the Weibull distribution, Wave 2 by the Log-Normal distribution, and Wave 3 by the Generalised Gamma distribution. These findings indicate that COVID-19 daily case patterns are asymmetric, heterogeneous, and wave-specific. Hence, flexible positive-support and right-skewed distributions are more useful for describing the empirical distribution of positive daily confirmed COVID-19 cases than symmetric assumptions. However, the selected distributions are interpreted only as relative empirical approximations among the evaluated candidate distributions, and not as exact distributional laws of COVID-19 daily cases.

Keywords: COVID-19, Distribution pattern analysis, Daily confirmed cases, Parametric distribution fitting, Skewness, Kurtosis, India epidemic waves, Public health statistics.

1. INTRODUCTION

COVID-19 has shown highly heterogeneous epidemic trajectories across different countries. The timing of the outbreak, the intensity of spread, the peak size, the duration of waves, and the reporting pattern varied substantially from one country to another. Due to this heterogeneity, the statistical analysis of daily confirmed cases is important not only to describe the epidemic burden but also to understand how cases are distributed over time [3]. A cross-country comparison of COVID-19 is useful, as countries differ in population size, health-care systems, public health responses and epidemic timing [1, 2]. These factors may result in different observed patterns in reported daily cases [2, 3].

In this study, the distribution pattern of COVID-19 daily cases is carried out for six countries: India, the United States, Brazil, the United Kingdom, South Africa, and Japan. The daily confirmed COVID-19 case data are taken from the Our World in Data COVID-19 dataset, which provides internationally comparable time-series data compiled from official sources [4]. Our World in Data describes confirmed cases as people tested and confirmed to be infected with SARS-CoV-2, but also notes that not everyone is tested, so confirmed cases are lower than the true number of infections [4]. The study period is restricted to a common analytical window from 30 January 2020 to 20 February 2022. After preprocessing, 753 daily observations are obtained for each country. A common observation period is used so that the distributional comparisons are not affected by unequal time coverage across the selected countries.

The primary objective of this study is to analyse the distributional behaviour of daily COVID-19 cases across different countries and across epidemic waves in India. This analysis is important because epidemic time series are generally not symmetric and do not follow a normal distribution. Daily case counts are non-negative, often right-skewed, and may contain extreme peak values during the period of rapid transmission. Hence, the mean value alone is not sufficient to describe the behaviour of daily cases. Distributional characteristics such as median, variance, skewness, and kurtosis provide additional information about asymmetry, dispersion, and tail behaviour.

*Address correspondence to this author at the Department of Statistics, School of Applied Sciences, Amity University, Jaipur, Rajasthan 303002, India;
E-mail: sahu01anita@gmail.com

These characteristics are particularly important in the case of COVID-19, as epidemic waves usually show rapid growth, a sharp peak, and an uneven declining phase.

For India, the analysis is further divided into three epidemic waves based on the classification used by Kumar *et al*. [5]. This wave-wise analysis is important because the complete Indian COVID-19 time series cannot be considered as a single homogeneous distribution. Treating the full series as one distribution may hide the differences between different epidemic phases. Hence, wave-specific analysis helps to examine whether the first, second, and third waves in India differ not only in magnitude but also in distributional shape, skewness, and extremity.

Accordingly, this study uses a distributional statistical perspective. The distributional analysis is used to describe the shape of daily cases across different countries and Indian epidemic waves. This analysis helps to go beyond the simple reporting of total cases and allows the study to analyse the empirical structure of daily COVID-19 case dynamics.

The present study is limited to distribution pattern analysis of reported daily confirmed COVID-19 cases. The fitted distributions are used to describe the empirical distributional structure of the observed daily case series. They are not used as epidemic transmission models or as independent data-generating laws. This distinction is important because COVID-19 daily case observations are time-ordered and may be affected by temporal dependence, reporting practices, testing behaviour, interventions, and changing epidemic phases

2. MATERIALS AND METHODS

2.1. Data Source and Study Design

In this study, daily confirmed COVID-19 case data are taken from the Our World in Data COVID-19 dataset. This dataset provides internationally comparable COVID-19 time-series data in country-date format, where each country has one observation for each date. It has been widely used for global pandemic monitoring and comparative analysis [4]. Confirmed case counts are selected because they provide a direct daily measure of reported epidemic activity and are consistently available for the selected countries during the study period. However, confirmed cases do not represent the total number of infections, as they depend on testing availability, testing behaviour, reporting system, and national surveillance practices [4, 6, 19]. Hence, the results of this study are interpreted as patterns observed in reported daily confirmed cases, and not as complete estimates of all SARS-CoV-2 infections.

Although per-million scaling is used for cross-country comparison, this adjustment only accounts for population size. It does not adjust for differences in testing rates, testing eligibility, case definitions, surveillance capacity, or national reporting practices. Therefore, the country-wise distributional differences observed in this study may reflect both epidemic intensity and reporting or surveillance differences across countries.

The analysis is carried out on six countries: India, the United States, Brazil, the United Kingdom, South Africa, and Japan. These countries are selected to include variation in epidemic burden, geographical context, health-system setting, and observed pandemic trajectory. Cross-country comparison is suitable in this study because COVID-19 case patterns varied substantially across different national settings. Differences in public-health conditions, surveillance systems, and policy contexts may contribute to heterogeneous pandemic outcomes across countries, as observed in recent comparative studies [2, 16]. India is considered the central country of interest in this analysis. This allows the study to compare India externally with other selected countries and internally across different epidemic waves.

For comparability, all country-level time series are harmonised to a common analytical window from 30 January 2020 to 20 February 2022. As a result, 753 daily observations are obtained for each country. This common window is used because unequal observation

Table 1: Harmonised Daily COVID-19 Case Data for Selected Countries

Country	Rows	Start	End	Imputed days	Mean new cases	Maximum new cases
India	753	2020-01-30	2022-02-20	0	56,869.154	414,188
United States	753	2020-01-30	2022-02-20	0	103,420.781	1,265,520
Brazil	753	2020-01-30	2022-02-20	1	37,292.979	298,408
United Kingdom	753	2020-01-30	2022-02-20	0	24,806.547	275,647
South Africa	753	2020-01-30	2022-02-20	0	4,856.695	37,875
Japan	753	2020-01-30	2022-02-20	0	5,955.838	103,038

Table 2: Descriptive Summary of Daily Confirmed COVID-19 Cases Across Indian Epidemic Waves

Wave	Rows	Start	End	Total cases	Mean	Median	Std. dev.	Min	Max	Skewness	Kurtosis
Wave 1	369	2020-01-30	2021-02-01	10,757,610	29,153.415	19,148	28,575.915	0	97,894	0.747	-0.643
Wave 2	317	2021-02-02	2021-12-15	23,953,018	75,561.571	35,871	102,926.858	5,784	414,188	1.991	2.778
Wave 3	67	2021-12-16	2022-02-20	8,111,845	121,072.313	71,365	112,009.489	5,326	347,254	0.553	-1.188

periods may bias the distributional comparison by including different epidemic phases in different countries. After preprocessing, each country has the same number of daily observations. One missing calendar observation is identified for Brazil and imputed during preprocessing. No duplicate dates, missing daily case values, or negative daily case counts are found in the final analytical dataset.

2.2. Country Selection and Analytical Period

The six selected countries are India, the United States, Brazil, the United Kingdom, South Africa, and Japan. These countries are included to represent different geographical settings, population sizes, health-system contexts, and pandemic trajectories. The selected countries also provide a suitable basis for comparing daily confirmed case patterns after population standardisation.

The analytical period is restricted to 30 January 2020 to 20 February 2022. This period provides a common observation window for all selected countries. The same number of observations is used for each country so that the distributional comparison is not affected by unequal time coverage. After harmonisation, each country has 753 daily observations.

For country-level distributional comparison, daily cases per million population are used instead of raw daily cases. This is necessary because the selected countries differ largely in population size. If raw case counts are used directly, the comparison may partly reflect population scale rather than epidemic intensity. Per-million scaling provides a more meaningful comparison of reported incidence patterns across countries. However, this scaling does not remove the influence of country-level differences in testing intensity, case detection, and reporting systems. Raw daily case counts are still useful for describing national burden, but per-million values are more suitable for cross-country distributional modelling and visual comparison.

2.3. India Epidemic Wave Classification

For India-specific analysis, the data are divided into three epidemic waves based on the wave classification used by Kumar *et al*. [5]. Wave 1 includes the period from 30 January 2020 to 1 February 2021, Wave 2 includes the period from 2 February 2021 to 15 December 2021, and Wave 3 includes the period from 16 December 2021 to 20 February 2022.

This wave-wise division is important because India's COVID-19 trajectory was not homogeneous over time. Each wave had a different duration, peak intensity, and distributional shape. Hence, modelling the complete Indian series as a single distribution may hide the phase-specific patterns. Additional genomic and clinical studies also suggest that later Indian waves reflected changing variant contexts, especially during the Delta and Omicron periods [14, 15].

2.4. Data Preparation and Preprocessing

The raw COVID-19 time-series data are first filtered to include six countries: India, the United States, Brazil, the United Kingdom, South Africa, and Japan. The final analytical period is restricted to a common window from 30 January 2020 to 20 February 2022. As a result, a harmonised country-level dataset with 753 daily observations for each country is obtained. This harmonisation is necessary because distributional comparisons may be biased when countries are observed for unequal time periods or during different epidemic phases.

The main variable used in this study is daily confirmed new cases. Confirmed case data provide a standardised daily measure of reported epidemic activity and are available consistently across the selected countries in the Our World in Data COVID-19 dataset. However, confirmed cases should not be considered as complete infection counts. They depend on testing availability, reporting practices, surveillance capacity, and national case definitions. Our World in Data defines confirmed cases as people who tested positive for SARS-CoV-2, and the true number of infections may be higher because not all infected individuals are tested [4]. Similar limitations are also reported in public-health studies, where under-reporting and variation in reporting systems are considered important limitations in cross-country comparison of COVID-19 case data [6, 16, 19]. Hence, all findings in this study are interpreted as patterns in reported confirmed COVID-19 cases, and not as estimates of total SARS-CoV-2 infections.

The preprocessing procedure includes checking duplicate dates, missing calendar dates, missing daily

case values, and negative daily case counts. The final prepared dataset contains no duplicate dates, no missing case values, and no negative daily case counts. One missing calendar observation is found for Brazil and is imputed using linear interpolation. Since this represents only one day out of 753 observations, it is unlikely to have a major effect on the distributional comparison. Negative values, if present, are constrained to zero as a data-quality safeguard, because negative daily case counts are not meaningful for the distributional modelling used in this study.

For cross-country comparison, daily cases are converted into daily cases per million population using 2022 population denominators. This transformation is necessary because the selected countries differ substantially in population size. Raw daily case counts describe national case burden, but they are not directly comparable across countries. Per-million scaling provides a more suitable basis for comparing reported epidemic intensity across countries with different population sizes.

For India-specific analysis, raw daily case counts are retained because the comparison is within a single country. Since the population denominator remains constant across Indian waves, raw counts are appropriate for comparing wave intensity, peak magnitude, and distributional structure. The Indian time series is divided into three epidemic waves using the classification based on Kumar *et al*. [5]: Wave 1 from 30 January 2020 to 1 February 2021, Wave 2 from 2 February 2021 to 15 December 2021, and Wave 3 from 16 December 2021 to 20 February 2022. This wave-wise segmentation is necessary because India's COVID-19 trajectory was not homogeneous over time. Each wave differed in timing, duration, peak intensity, and distributional form.

For descriptive analysis, zero daily case values are retained because they are valid observations in the reported case series. However, for parametric distribution fitting, the fitting sample is restricted to positive daily case observations. This common positive-case fitting sample is used for all candidate distributions, including the Normal reference model, so that AIC, BIC, RMSE, and MAE can be compared across models fitted on the same observations. This step is required because most candidate distributions considered in this study, such as Gamma, Generalised Gamma, Weibull, Log-Normal, Log-Logistic, Inverse Gaussian, and Exponential, are continuous positive-support distributions. The same positive-case sample is also used for the Normal reference model to maintain fair model comparison across all candidate distributions. Therefore, the fitted distributions should be interpreted as conditional distributions of daily confirmed cases given that reported cases were positive. This positive-case fitting approach may introduce bias if zero-case days are epidemiologically meaningful or if the proportion of zero observations is not negligible. This issue is especially relevant for Brazil and South Africa, where approximately 5% of observations are excluded from fitting. Zero-inflated or hurdle models may provide a more complete approach for modelling zero and positive observations together, but they are outside the scope of the present distributional comparison.

3. STATISTICAL METHODS

3.1. Descriptive Distributional Analysis

Descriptive distributional analysis is performed before fitting the parametric distributions. This step is important because daily COVID-19 case series may differ not only in total number of cases or peak size, but also in mean, median, variation, skewness and tail behaviour. These measures give the first statistical view of daily cases before applying the distribution fitting methods.

For cross-country analysis, descriptive statistics are calculated using both raw daily cases and daily cases per million population. Daily cases per million are more

Table 3: Data Preprocessing Summary

Step	Procedure	Rationale
Country filtering	India, United States, Brazil, United Kingdom, South Africa and Japan	Supports cross-country comparison
Date harmonisation	30 Jan 2020–20 Feb 2022	Ensures equal observation window
Calendar check	Daily sequence checked for missing dates	Preserves time-series continuity
Imputation	One Brazil date imputed using linear interpolation	Maintains complete country panel
Negative-value check	Values constrained to zero if negative	Ensures valid non-negative case series
Per-million scaling	Used for country-level comparison	Adjusts for population size
India wave segmentation	Three waves based on Kumar *et al*. [5]	Enables phase-specific analysis
Zero handling	Zeros retained descriptively; positive-only sample used for distribution fitting	Ensures valid comparison among positive-support distributions; fitted models are interpreted as conditional on positive reported cases

useful for comparing countries because the selected countries have different population sizes. Raw daily cases are also reported because they show the actual national burden of reported COVID-19 cases.

To avoid depending only on mean, skewness and kurtosis, robust distributional measures are also calculated. These include interquartile range, median absolute deviation, 90th percentile, 95th percentile and max-to-median ratio. These measures are useful because they are less affected by extreme values compared with skewness and kurtosis.

For India-specific analysis, descriptive statistics are calculated separately for Wave 1, Wave 2 and Wave 3. This is required because the complete Indian case series contains different epidemic phases. If all waves are combined, the internal differences between waves may not be clearly identified.

3.2. Candidate Distributions

The primary objective of the distributional analysis is to identify how daily COVID-19 cases are distributed across different countries and across epidemic waves in India. Since daily case counts are non-negative, temporally clustered, and generally asymmetric, different candidate distributions are considered for the analysis. These distributions are selected in order to capture different levels of skewness, tail behaviour, and distributional flexibility. The purpose of this analysis is not to state that any single distribution is the true epidemic-generating distribution. Rather, the objective is to identify the best-fitting distributional approximation from a predefined set of candidate distributions.

In this work, eight candidate distributions are evaluated: Gamma, Generalised Gamma, Weibull, Log-Normal, Log-Logistic, Inverse Gaussian, Exponential, and Normal. These distributions are selected because they include positive-support, skewed, flexible, and benchmark distributions. Gamma, Weibull, and Log-Normal distributions are often used in infectious-disease modelling, especially for positive epidemiological variables such as incubation periods, serial intervals, and reporting delays [8, 10, 12]. Although daily case counts are not the same as delay distributions, they have similar statistical characteristics such as non-negativity, right-skewness, and possible long-tail behaviour.

The Gamma distribution is included because it is a flexible positive-support distribution and can represent right-skewed data with different levels of dispersion. Hence, it is suitable for epidemic case patterns where most days have moderate case counts, while some days have very high case counts. The Weibull distribution is included because its shape parameter allows it to represent different types of tail behaviour and wave-like patterns. The Log-Normal distribution is included because epidemic growth and biological processes may sometimes follow multiplicative mechanisms, where proportional changes are more important than additive changes. The Log-Logistic distribution is included as a heavier-tailed alternative to the Log-Normal distribution, so that the analysis can examine whether extremely high-case days are better represented by a stronger tail distribution.

The Generalised Gamma distribution is included because it is a flexible distributional family and can represent a wider range of shapes than the standard Gamma or Weibull distributions. Its inclusion is useful because different epidemic waves may differ in skewness, spread, and tail behaviour. The Inverse Gaussian distribution is included as another positive and right-skewed candidate distribution. It is treated as a secondary candidate distribution because it provides an alternative shape for asymmetric positive data, but it is less central than Gamma, Weibull, Log-Normal, and Generalised Gamma in epidemiological distributional modelling. The Exponential distribution is retained as a simple benchmark model. It is restrictive and may not capture the full structure of epidemic waves, but it provides a parsimonious reference for positive-valued data. The Normal distribution is included as a

Table 4: Candidate Distributions and Rationale

Distribution	Role in analysis	Rationale
Gamma	Core candidate	Flexible positive right-skewed distribution
Weibull	Core candidate	Captures varied skewness and tail forms
Log-Normal	Core candidate	Suitable for multiplicative positive processes
Log-Logistic	Core candidate	Heavy-tailed positive alternative
Generalized Gamma	Core candidate	Flexible family for varied shapes
Inverse Gaussian	Secondary candidate	Additional positive right-skewed form
Exponential	Benchmark	Simple positive-support reference model
Normal	Reference	Symmetric benchmark for comparison

symmetric reference model. It is not expected to fit daily COVID-19 cases well because daily cases are non-negative and generally skewed. However, including the Normal distribution helps to check whether symmetric assumptions are empirically suitable when compared with flexible asymmetric alternatives.

3.3. Distribution Fitting and Model Comparison

For descriptive statistics, zero daily case values are retained because they are valid observations in the reported case series. However, for parametric distribution fitting, the fitting sample is restricted to positive daily case observations because most of the candidate distributions used in this study are continuous positive-support distributions. This means that the fitted models describe the conditional distribution of daily reported cases given that reported cases are positive. To make this step clear, the number and percentage of zero observations excluded from each fitting sample are reported before the distribution-fitting results. The same positive-case fitting sample is used for all candidate distributions, including the Normal reference model, so that AIC, BIC, RMSE, and MAE are compared across models fitted on the same observations.

For cross-country distributional analysis, daily cases per million population are used. This adjustment is necessary because the selected countries differ substantially in population size, and raw daily case counts may reflect population scale rather than comparable epidemic intensity. For India's wave-wise analysis, raw daily case counts are retained because the comparison is within the same country, and the population denominator remains constant across waves.

Each candidate distribution is fitted using maximum likelihood estimation. Maximum likelihood estimation is a standard method for estimating distribution parameters by selecting the parameter values that make the observed data most probable under the assumed distribution [20]. After fitting the candidate distributions, the models are compared using information criteria and empirical fit measures. The Akaike Information Criterion is used because it balances goodness-of-fit with model complexity. The Bayesian Information Criterion is also reported as an additional complexity-penalised measure [21]. AIC is widely used for model selection because it evaluates relative model quality while penalising unnecessary parameters [7, 17].

The final distribution ranking is based on a composite score obtained by combining the ranks of AIC, RMSE, and MAE. Equal weighting is used for AIC, RMSE, and MAE because no single criterion is considered sufficient to represent all aspects of fit, and no prior empirical basis is available for assigning greater weight to one criterion over another. Therefore, equal weighting is used as a transparent and neutral ranking approach. To avoid overdependence on this analytical choice, sensitivity analysis is also performed using AIC-only, BIC-only, RMSE-only, and MAE-only rankings. This approach is used because likelihood-based criteria and empirical density-error measures do not always measure the same aspect of model fit. AIC evaluates relative likelihood with a penalty for model complexity, whereas RMSE and MAE compare the fitted density with the empirical density approximation. In this study, RMSE and MAE are calculated by comparing the fitted probability density with the empirical histogram density using the same binning structure within each dataset. Since the combination of these criteria involves an analytical choice, sensitivity analysis is also performed using AIC-only, BIC-only, RMSE-only, and MAE-only rankings. Where the composite ranking differs from AIC-only or BIC-only selection, the result is interpreted cautiously as model-selection uncertainty. BIC, KS statistic, and KS p-value are reported as additional diagnostic measures, but they are not used as the only basis for distribution selection.

Table 5: Model-Selection and Diagnostic Criteria Used in the Distribution-Fitting Analysis. AIC, RMSE, and MAE are Included in the Composite Ranking, Whereas BIC, KS Statistic, Standard KS P-Value, and Bootstrap KS P-Value are Reported As Supplementary Diagnostic Measures

Criterion	Purpose	Use in this study
AIC	Relative model fit with complexity penalty	Included in equal-weight composite ranking
BIC	Stronger complexity penalty	Reported as supplementary criterion
RMSE	Density approximation error	Included in equal-weight composite ranking
MAE	Average density approximation error	Included in equal-weight composite ranking
KS statistic	CDF discrepancy	Diagnostic measure
KS p-value	Approximate goodness-of-fit indicator	Interpreted cautiously
Bootstrap KS p-value	Goodness-of-fit diagnostic accounting for parameter estimation through simulation and refitting	Reported for selected best-fitting distributions

Table 6: Country-Wise Descriptive Statistics of Daily COVID-19 Cases

Country	Population 2022	Rows	Total cases	Mean per million	Median per million	SD per million	Mean	Median	SD	Min	Max	Skewness	Kurtosis
India	1,417,173,173	753	42,822,473.0	40.129	20.744	58.375	56,869.154	29,398.0	82,727.941	0.0	414,188.0	2.525	5.989
United States	338,289,857	753	77,875,848.0	305.716	176.145	425.447	103,420.781	59,588.0	143,924.539	0.0	1,265,520.0	3.902	18.902
Brazil	215,313,498	753	28,081,613.5	173.203	132.514	174.440	37,292.979	28,532.0	37,559.285	0.0	298,408.0	2.369	8.862
United Kingdom	67,508,936	753	18,679,330.0	367.456	176.895	527.713	24,806.547	11,942.0	35,625.312	0.0	275,647.0	3.129	13.531
South Africa	59,893,885	753	3,657,091.0	81.088	38.852	94.572	4,856.695	2,327.0	5,664.270	0.0	37,875.0	1.620	2.638
Japan	123,951,692	753	4,484,746.0	48.050	8.657	132.431	5,955.838	1,073.0	16,415.070	0.0	103,038.0	4.331	18.874

The Kolmogorov–Smirnov statistic is used to measure the discrepancy between the empirical and fitted cumulative distribution functions [11, 13]. Since the standard KS p-value is not exact when distribution parameters are estimated from the same data, a parametric bootstrap KS procedure is also used for the selected best-fitting distributions [9, 13, 18]. For each selected distribution, data of the same sample size are repeatedly simulated from the fitted model using B = 500 bootstrap replications. The distribution is then refitted to each simulated sample, and the KS statistic is recalculated [9, 18]. The bootstrap p-value is calculated as the proportion of simulated KS statistics greater than or equal to the observed KS statistic [9]. Hence, the KS results are interpreted as diagnostic evidence of relative fit, and not as proof that the data follow a particular parametric distribution [13, 18].

This methodology supports a cautious and comparative interpretation of distributional fitting. A selected distribution is described only as the best-fitting candidate among the evaluated distributions, and not as the true distribution of COVID-19 cases. This distinction is important because epidemic case data are temporally dependent, affected by reporting systems, and influenced by external factors such as interventions, variants, and testing practices. Therefore, the distributional analysis is used to describe empirical distributional patterns, and not to infer a fixed biological law of epidemic spread.

Since daily COVID-19 case observations are time-ordered, they may show serial dependence and temporal clustering. The present distribution fitting does not model autocorrelation or epidemic transmission dynamics. Hence, the fitted distributions are interpreted as empirical marginal approximations of daily reported case values over the study period, rather than as independent data-generating models. This limitation is considered while interpreting the goodness-of-fit results.

4. RESULTS

4.1. Country-wise Descriptive Distributional Analysis

Daily case distributions are right-skewed in all six countries. In all countries, the mean value is greater than the median value. This indicates that the average daily cases are affected by some high-incidence days. Such a pattern is common in epidemic waves, where a few peak days may strongly increase the mean value.

Japan has the highest skewness value, followed by the United States and the United Kingdom. This shows that these countries have more asymmetric daily case distributions during the study period. South Africa has the lowest skewness among the selected countries, but its distribution is also positively skewed. Hence, none of the selected countries shows a symmetric distribution of daily cases.

Kurtosis values also show large differences across countries. The United States and Japan have the highest kurtosis values, which indicates the presence of extremely high-case observations. The United Kingdom also shows high kurtosis. South Africa has lower kurtosis, which indicates comparatively less extreme tail behaviour. These results show that the daily case distribution is not explained only by average incidence. The countries also differ in how much their case series are affected by extreme peak days.

Per-million statistics give a different interpretation than raw case counts. The United States has the largest total number of cases in the dataset. However, the United Kingdom has the highest mean daily cases per million, followed by the United States and Brazil. India has a lower mean cases per million, although its total number of cases is large. This shows that population adjustment is necessary for cross-country comparison. Japan has low median cases per million but high skewness and kurtosis, which indicates that its distribution contains long low-incidence periods and short high-incidence periods.

Figure **1** visually supports the descriptive statistics reported in Table 6. The compression of the boxes near the lower range and the presence of many upper-tail observations confirm that daily case distributions were highly asymmetric. This reinforces the need to use distributional measures such as skewness and kurtosis, and later to evaluate flexible positive-support distributions rather than relying on symmetric assumptions.

Figure **2** complements the boxplot by showing that the observed distributional differences are generated by dynamic epidemic trajectories rather than by static differences in average case levels alone. Countries with sharp or late epidemic peaks tend to show stronger right-skewness and heavier upper tails, while countries with repeated moderate waves show different dispersion patterns. Therefore, the time-series plots provide important context for interpreting the descriptive statistics and the subsequent distribution-fitting results.

4.2. India Wave-wise Descriptive Distributional Analysis

For India-specific analysis, descriptive statistics are calculated separately for Wave 1, Wave 2 and Wave 3.

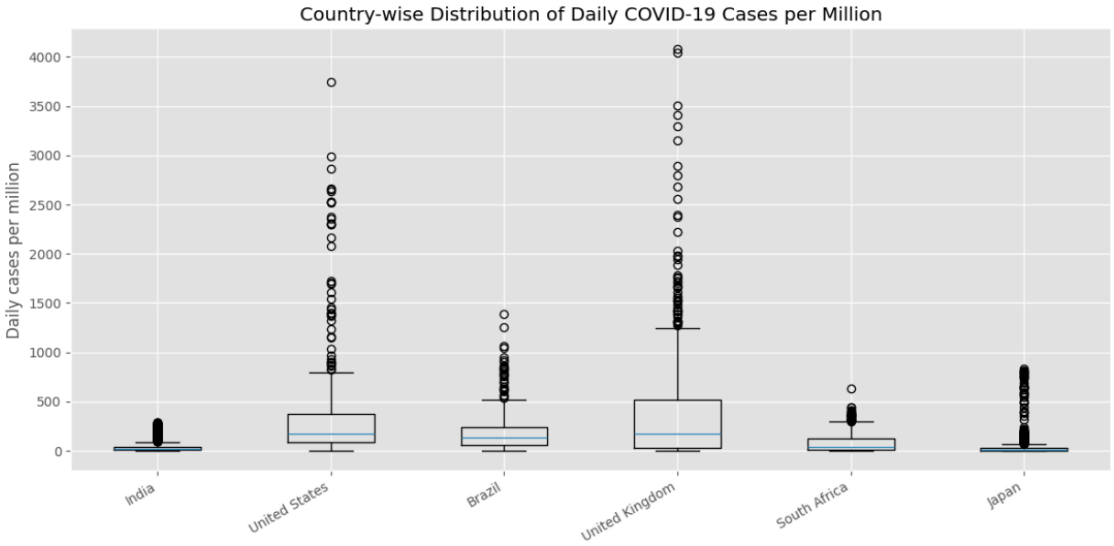

Figure 1: Country-wise distribution of daily COVID-19 cases per million, 30 January 2020 to 20 February 2022.

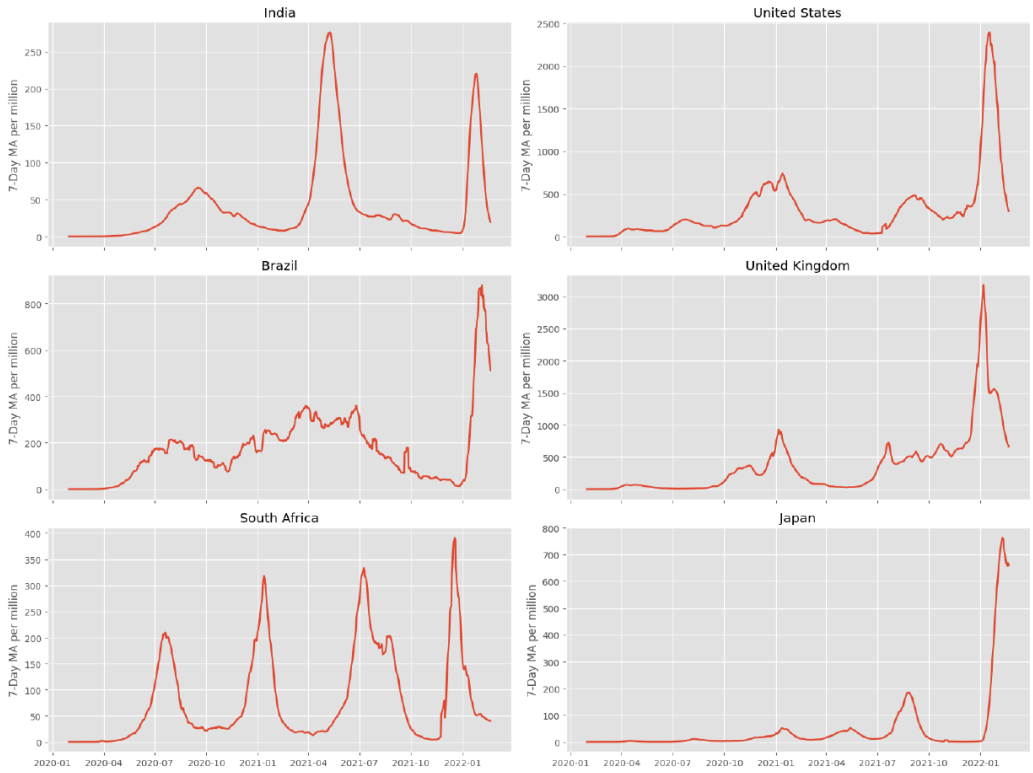

Figure 2: Country-wise time series of 7-day moving average daily COVID-19 cases per million, 30 January 2020 to 20 February 2022.

This is required because the complete Indian case series contains different epidemic phases. If all waves are combined, the internal differences between waves may not be clearly identified.

The results show that the three waves are different in both magnitude and distributional shape. Wave 1 has a lower mean and median daily cases compared with Wave 2 and Wave 3. It has moderate positive skewness and negative kurtosis, which indicates that the distribution is less extreme compared with later waves.

Wave 2 has much higher mean daily cases and the highest maximum daily case count. It also has strong positive skewness and positive kurtosis. The difference between mean and median is also large in Wave 2. This indicates that Wave 2 is strongly affected by high-incidence peak days. Hence, Wave 2 is not only larger in total burden but also more extreme in distributional form.

Wave 3 has the highest mean and median daily cases, but it contains only 67 days. Therefore, the interpretation of Wave 3 should be done carefully.

Since the wave period is short, its descriptive statistics may be more affected by the selected wave boundary and by individual high-incidence days. The negative kurtosis and lower skewness suggest that the distribution within this wave is less heavy-tailed than that of Wave 2.

Overall, the descriptive analysis shows that COVID-19 daily case distributions are asymmetric and heterogeneous across countries and across Indian waves. This conclusion is supported by mean-median differences, skewness, kurtosis and robust statistics. The results also justify the use of flexible positive-support distributions such as Gamma, Weibull, Log-Normal, Log-Logistic and Generalised Gamma, instead of relying only on symmetric models.

Figure **3** visually supports the descriptive results in Table **7** by showing that India's epidemic waves differed not only in total case burden but also in empirical distributional shape. These differences justify the use of wave-specific analysis rather than modelling India's full time series as a single homogeneous distribution.

Figure **4** provides a temporal context for the wave-wise descriptive analysis. The visible separation between the three epidemic phases supports the decision to analyse India by wave, since a full-period distribution would combine distinct phases with different peak structures and durations.

4.3. Positive-case Fitting Samples and Zero Exclusions

After descriptive analysis, parametric distribution fitting is performed to find which candidate distribution gives a better approximation to the daily COVID-19

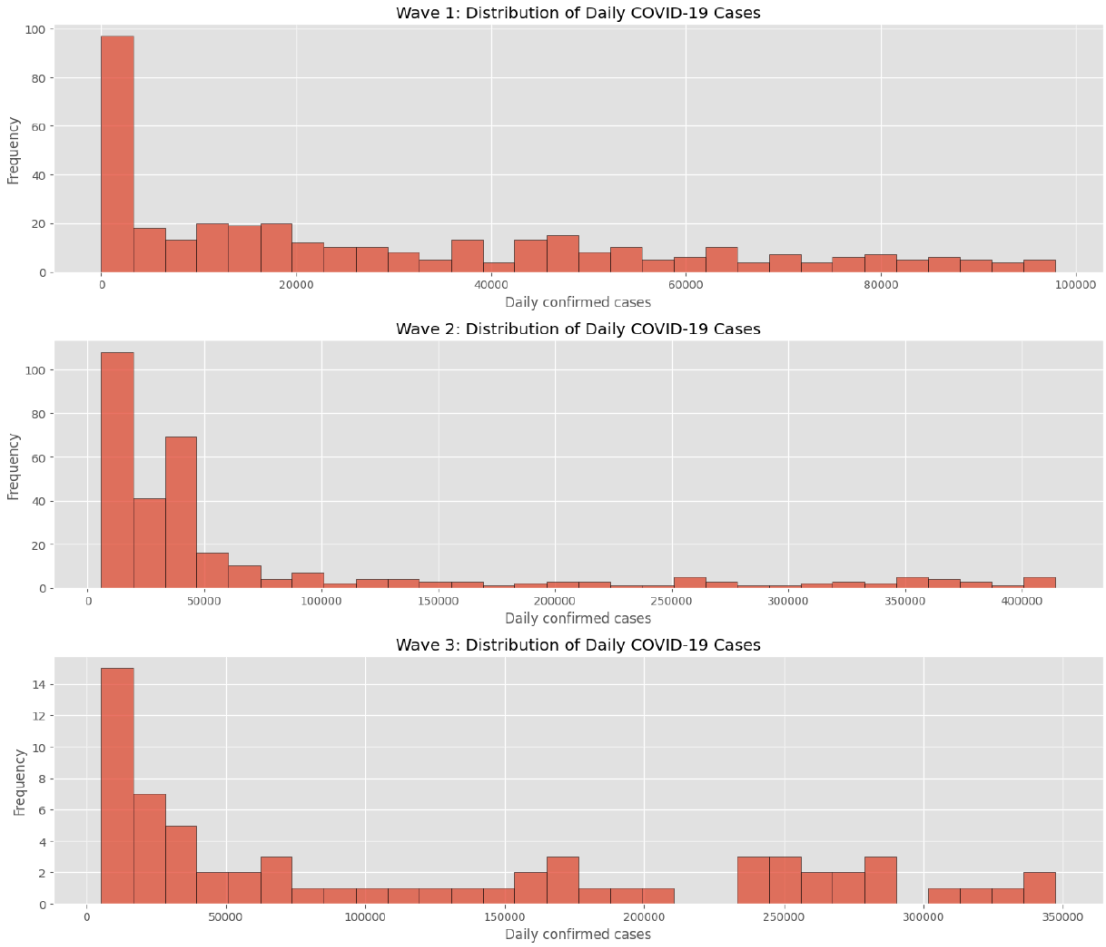

Figure 3: India wave-wise distribution of daily confirmed COVID-19 cases.

Table 7: India Wave-Wise Descriptive Statistics

Wave	Rows	Start	End	Total cases	Mean	Median	SD	Min	Max	Skewness	Kurtosis
Wave 1	369	2020-01-30	2021-02-01	10,757,610	29,153.415	19,148.0	28,575.915	0	97,894	0.747	-0.643
Wave 2	317	2021-02-02	2021-12-15	23,953,018	75,561.571	35,871.0	102,926.858	5,784	414,188	1.991	2.778
Wave 3	67	2021-12-16	2022-02-20	8,111,845	121,072.313	71,365.0	112,009.489	5,326	347,254	0.553	-1.188

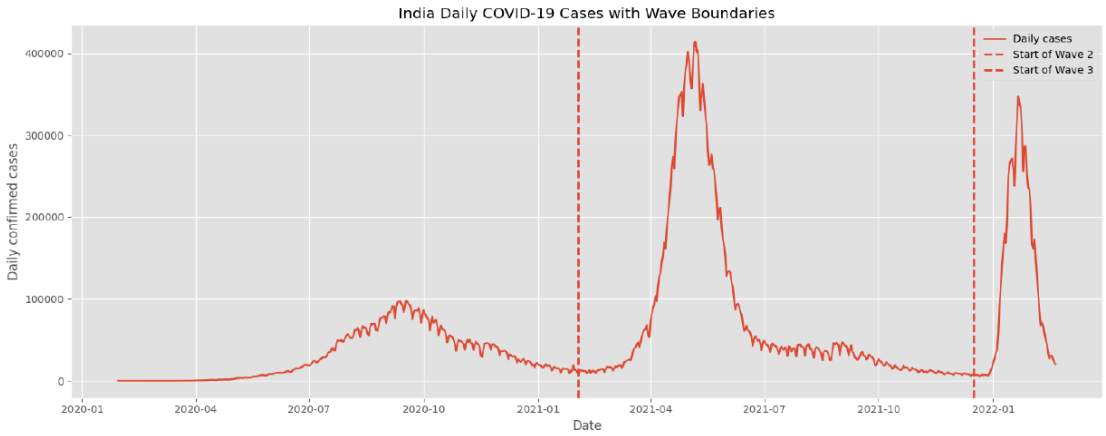

Figure 4: India daily confirmed COVID-19 cases with epidemic wave boundaries, 30 January 2020 to 20 February 2022.

case data. For country-wise analysis, the distributions are fitted on positive daily cases per million population. For India's wave-wise analysis, the distributions are fitted on positive raw daily case counts for each wave. The selected distribution is interpreted only as the best-fitting distribution among the candidate distributions used in this study. It is not considered the true distribution of COVID-19 cases.

Zero values are kept in descriptive analysis because they are valid reported observations. However, for distribution fitting, zero values are removed because most of the selected candidate distributions are defined only for positive values. Hence, the fitted distributions describe daily case behaviour only when the reported cases are positive.

From the positive-case fitting sample, it is observed that only a small number of zero observations are excluded at the country level. Japan has the lowest percentage of excluded zero observations, while Brazil and South Africa have the highest percentage among the selected countries. For the India wave-wise analysis, zero observations are excluded only in Wave 1. Wave 2 and Wave 3 have no zero daily case observations in the selected wave periods.

Although the percentage of excluded zero observations is relatively small, their exclusion may still affect the fitted distribution because zero-case days represent part of the reported epidemic pattern. This issue is more relevant for Brazil and South Africa, where 5.05% of observations are excluded, and for Indian Wave 1, where 8.13% of observations are excluded. Therefore, the fitted distributions should be interpreted as positive-case distributions and not as complete distributions of all reported daily observations.

4.4. Country-wise Distribution Fitting Results

The final distribution selection is based on the composite ranking of AIC, RMSE and MAE. KS statistic and KS p-value are used as diagnostic measures. This is necessary because the distribution with the lowest AIC may not always give the lowest density error. Hence, the selection is not based on a single criterion only.

Sensitivity analysis is also performed to check whether the selected distributions remain the same under different criteria. For this, the best distribution is separately identified using AIC-only, BIC-only, RMSE-only and MAE-only rankings. This analysis is important because likelihood-based criteria and density-error criteria may select different distributions. If different criteria select different distributions, then the result is interpreted with model-selection uncertainty.

Table **9** shows the sensitivity of selected distributions to alternative ranking criteria. From the table, it is observed that the composite ranking generally agrees more with RMSE-only and MAE-only rankings than with AIC-only or BIC-only rankings. For India, Japan and the United States, the Log-Logistic distribution is selected by composite ranking as well as by RMSE-only and MAE-only criteria. However, AIC and BIC select different distributions. This indicates that the selected distributions should be considered as the best empirical approximation under the selected composite criterion, not as the only possible distribution.

The results show that no single distribution is best for all countries. Log-Logistic distribution is selected for India, Japan and the United States. The inverse Gaussian distribution is selected for Brazil. The Generalised Gamma distribution is selected for South Africa. Log-Normal distribution is selected for the United Kingdom.

For Brazil, India, Japan, South Africa and the United States, the bootstrap KS p-values do not show strong disagreement between the fitted distribution and observed positive-case distribution. This means that

Table 8: Positive-Case Fitting Samples and Zero Exclusions for Distribution Fitting

Dataset	Analysis level	Total days	Positive days used	Zero days excluded	Zero days excluded (%)
India	Country full period per million	753	723	30	3.98
United States	Country full period per million	753	741	12	1.59
Brazil	Country full period per million	753	715	38	5.05
United Kingdom	Country full period per million	753	739	14	1.86
South Africa	Country full period per million	753	715	38	5.05
Japan	Country full period per million	753	747	6	0.80
Wave 1	India wave	369	339	30	8.13
Wave 2	India wave	317	317	0	0.00
Wave 3	India wave	67	67	0	0.00

Table 9: Sensitivity of Selected Distributions to Alternative Ranking Criteria

Dataset	Analysis level	Composite best	AIC-only best	BIC-only best	RMSE-only best	MAE-only best	Sensitivity interpretation
Brazil	Country	Inverse Gaussian	Exponential	Exponential	Inverse Gaussian	Inverse Gaussian	Composite selection differs from AIC and/or BIC; interpret with model-selection uncertainty
India	Country	Log-Logistic	Gamma	Gamma	Log-Logistic	Log-Logistic	Composite selection differs from AIC and/or BIC; interpret with model-selection uncertainty
Japan	Country	Log-Logistic	Generalized Gamma	Generalized Gamma	Log-Logistic	Log-Logistic	Composite selection differs from AIC and/or BIC; interpret with model-selection uncertainty
South Africa	Country	Generalized Gamma	Generalized Gamma	Generalized Gamma	Log-Logistic	Generalized Gamma	Composite selection agrees with AIC and BIC
United Kingdom	Country	Log-Normal	Generalized Gamma	Generalized Gamma	Log-Normal	Inverse Gaussian	Composite selection differs from AIC and/or BIC; interpret with model-selection uncertainty
United States	Country	Log-Logistic	Generalized Gamma	Generalized Gamma	Log-Logistic	Log-Logistic	Composite selection differs from AIC and/or BIC; interpret with model-selection uncertainty
Wave 1	India wave	Weibull	Gamma	Gamma	Weibull	Weibull	Composite selection differs from AIC and/or BIC; interpret with model-selection uncertainty
Wave 2	India wave	Log-Normal	Inverse Gaussian	Inverse Gaussian	Log-Normal	Log-Normal	Composite selection differs from AIC and/or BIC; interpret with model-selection uncertainty
Wave 3	India wave	Generalized Gamma	Gamma	Gamma	Generalized Gamma	Generalized Gamma	Composite selection differs from AIC and/or BIC; interpret with model-selection uncertainty

the selected distributions give useful empirical approximations for these countries. India and the United States show very high bootstrap KS p-values for the Log-Logistic distribution, which indicates that this distribution fits reasonably well under the bootstrap diagnostic.

South Africa needs careful interpretation. The standard KS p-value is below 0.05, but the bootstrap KS p-value is slightly above 0.05. Since the bootstrap procedure considers parameter estimation more properly, the result does not show a strong discrepancy at the 5% level. Hence, Generalised Gamma can be considered as a useful approximation for South Africa, but with caution.

The United Kingdom is the case where the selected distribution does not fully capture the observed distribution. Although Log-Normal is selected by the composite ranking, the bootstrap KS p-value is 0.000.

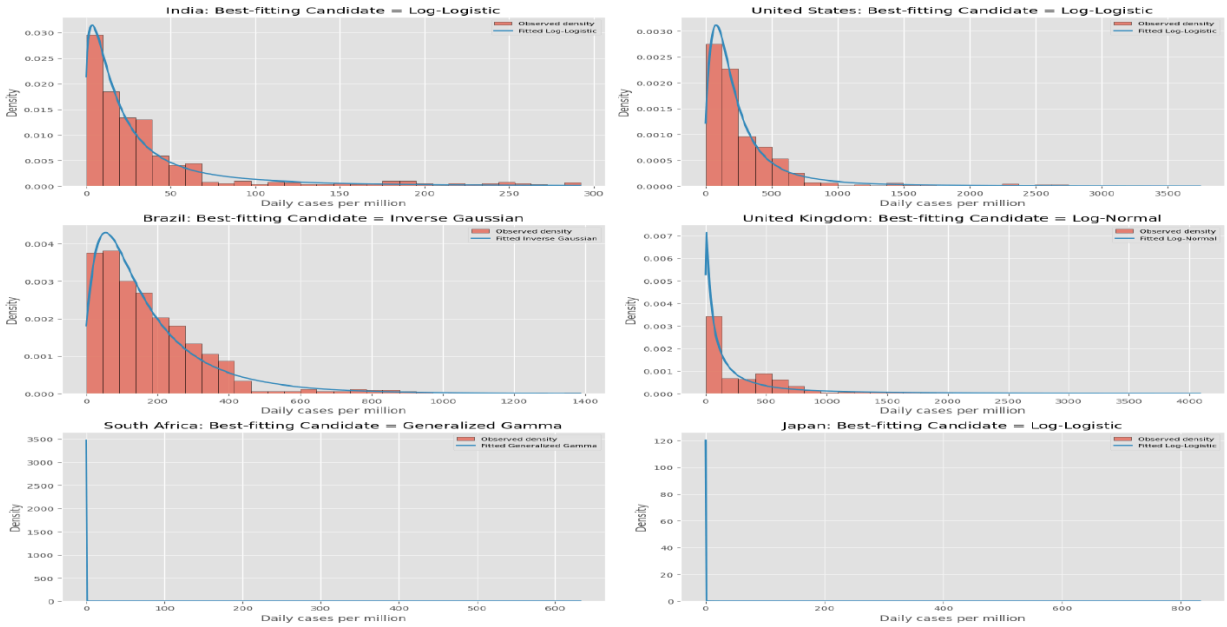

Figure 5: Country-wise fitted candidate distributions for positive daily COVID-19 cases per million.

This indicates that there is still a discrepancy between the fitted Log-Normal distribution and the empirical distribution. Hence, for the United Kingdom, the Log-Normal distribution is treated only as the best among the selected candidate distributions, not as an adequate distributional fit. Therefore, for the United Kingdom, no adequate distributional fit is identified among the evaluated candidate distributions.

Overall, the country-wise distribution fitting results show that flexible right-skewed distributions are more suitable for positive daily COVID-19 cases. Normal distribution performs poorly, which supports the descriptive result that daily cases are not symmetric. Exponential distribution is also generally weak under the composite ranking, although AIC and BIC sometimes favour simpler distributions. Hence, the result should not be interpreted as proof of one unique distributional form. It only shows that flexible asymmetric distributions give useful relative approximations for several observed positive-case datasets, although some datasets, especially the United Kingdom, still show inadequate fit under the bootstrap KS diagnostic. Table **10** shows the best-fitting country-wise distributions with bootstrap KS diagnostics.

The country-wise fitted curves show that flexible positive-support distributions generally captured the broad empirical shape of the observed case-per-million data better than symmetric alternatives. However, visual deviations remain in several panels, especially around tails and peak-density regions. This reinforces the interpretation that the selected models are best-fitting candidates within the evaluated set, not exact distributional laws.

4.5. India Wave-wise Distribution Fitting Results

For India wave-wise distribution fitting, the selected distribution changes across waves. Wave 1 is best approximated by the Weibull distribution. Wave 2 is best approximated by the Log-Normal distribution. Wave 3 is best approximated by the Generalised Gamma distribution. This shows that India's COVID-19 distributional pattern was not the same across different epidemic waves.

For Wave 1, the standard KS p-value is below 0.05, but the bootstrap KS p-value is 0.244. This indicates that after considering parameter estimation through bootstrap refitting, there is no strong evidence of discrepancy. Hence, the Weibull distribution can be considered as a reasonable empirical approximation for Wave 1.

For Wave 2, the Log-Normal distribution is selected by the composite ranking. However, the bootstrap KS p-value is approximately zero. This indicates a failed goodness-of-fit under the bootstrap KS diagnostic. Hence, the Log-Normal distribution should be interpreted only as the best relative candidate among the evaluated distributions, and not as an adequate parametric representation of Wave 2. This is also consistent with the descriptive analysis, where Wave 2 shows strong skewness, high kurtosis and extreme peak values.

For Wave 3, the Generalised Gamma distribution is selected, and the bootstrap KS p-value is 0.132. This indicates no strong bootstrap discrepancy. However, this result should be interpreted carefully because Wave 3 contains only 67 observations. Due to the small sample size, the fitted distribution may be

Table 10: Best-Fitting Country-Wise Distributions with Bootstrap KS Diagnostics

Country	Best-fitting candidate	AIC	BIC	KS statistic	Standard KS p-value	Bootstrap KS p-value	RMSE	MAE	N used	Bootstrap KS interpretation
Brazil	Inverse Gaussian	8933.416	8947.132	0.043821	0.124596	0.306	0.000182	0.000122	715	No strong bootstrap KS discrepancy
India	Log-Logistic	6790.899	6804.649	0.044431	0.111738	0.992	0.001134	0.000672	723	No strong bootstrap KS discrepancy
Japan	Log-Logistic	6168.024	6181.872	0.041590	0.146684	0.110	0.002314	0.000555	747	No strong bootstrap KS discrepancy
South Africa	Generalized Gamma	7684.925	7703.214	0.051635	0.042635	0.056	0.000778	0.000388	715	No strong bootstrap KS discrepancy
United Kingdom	Log-Normal	10121.913	10135.729	0.114228	<0.001	0.000	0.000154	0.000072	739	Bootstrap KS indicates inadequate fit; no adequate candidate distribution identified
United States	Log-Logistic	9950.766	9964.590	0.031981	0.425803	1.000	0.000081	0.000046	741	No strong bootstrap KS discrepancy

affected by the wave boundary and individual high-case days. Since the Generalised Gamma distribution has greater parameter flexibility, the stability of this fitted distribution may be more sensitive to the limited sample size. Therefore, the Wave 3 result is interpreted as exploratory and descriptive rather than definitive.

Figure **6** shows that the fitted distributions capture broad differences in wave-wise distributional form, but they should not be interpreted as evidence of exact model adequacy. This is particularly important for Wave 2, where the selected Log-Normal distribution was the best candidate under the composite ranking, but the bootstrap KS diagnostic indicated inadequate fit. For Wave 1 and Wave 3, bootstrap KS diagnostics did not indicate a strong discrepancy, although the fits should still be interpreted as empirical approximations rather than true generative models. The Wave 3 fit should also be interpreted cautiously because it is based on only 67 observations.

4.6. Summary of Distribution Fitting Findings

In summary, the distribution fitting results show four main points. First, positive daily COVID-19 cases are better represented by flexible asymmetric distributions than by Normal or simple Exponential distributions. Second, the best-fitting distribution differs across countries, which indicates heterogeneous case patterns even after population adjustment. Third, for India, the best-fitting distribution changes across waves, which supports wave-wise analysis. Fourth, sensitivity analysis and bootstrap KS diagnostics show that the selected distribution is not always stable across all criteria. Hence, all fitted distributions are interpreted as relative empirical approximations among the candidate distributions used in this study.

5. DISCUSSION

In this study, we analysed the distribution pattern of daily confirmed COVID-19 cases. The analysis is performed for six countries and also for three epidemic waves in India. The main purpose of this discussion is to combine the results of descriptive analysis and distribution fitting, and to explain what these results indicate about COVID-19 daily case behaviour.

The results show that COVID-19 daily case data are not homogeneous across countries. The descriptive analysis shows that all selected countries have right-skewed case distributions. In all countries, the mean value is higher than the median value, which indicates that the average case count is affected by high-incidence days. This is expected in epidemic data because during peak periods, the number of daily cases increases rapidly and produces extreme values. Hence, only the mean or total case count is not sufficient to explain the case pattern. Skewness, kurtosis and robust statistics are also required to understand the distributional behaviour.

The country-wise distribution fitting also supports this result. The same distribution is not selected for all countries. Log-Logistic distribution is selected for India,

Table 11: Best-Fitting Distributions across Indian Waves with Bootstrap KS Diagnostics

Wave	Best-fitting candidate	AIC	BIC	KS statistic	Standard KS p-value	Bootstrap KS p-value	RMSE	MAE	N used	Bootstrap KS interpretation
Wave 1	Weibull	7553.990	7565.468	0.130640	0.000017	0.244	0.000005	0.000004	339	No strong bootstrap KS discrepancy
Wave 2	Log-Normal	7606.320	7617.597	0.085601	0.018060	~0	0.000002	0.000001	317	Bootstrap KS indicates inadequate fit; best relative candidate only
Wave 3	Generalized Gamma	1702.508	1711.327	0.151551	0.082766	0.132	0.000002	0.000001	67	No strong bootstrap KS discrepancy

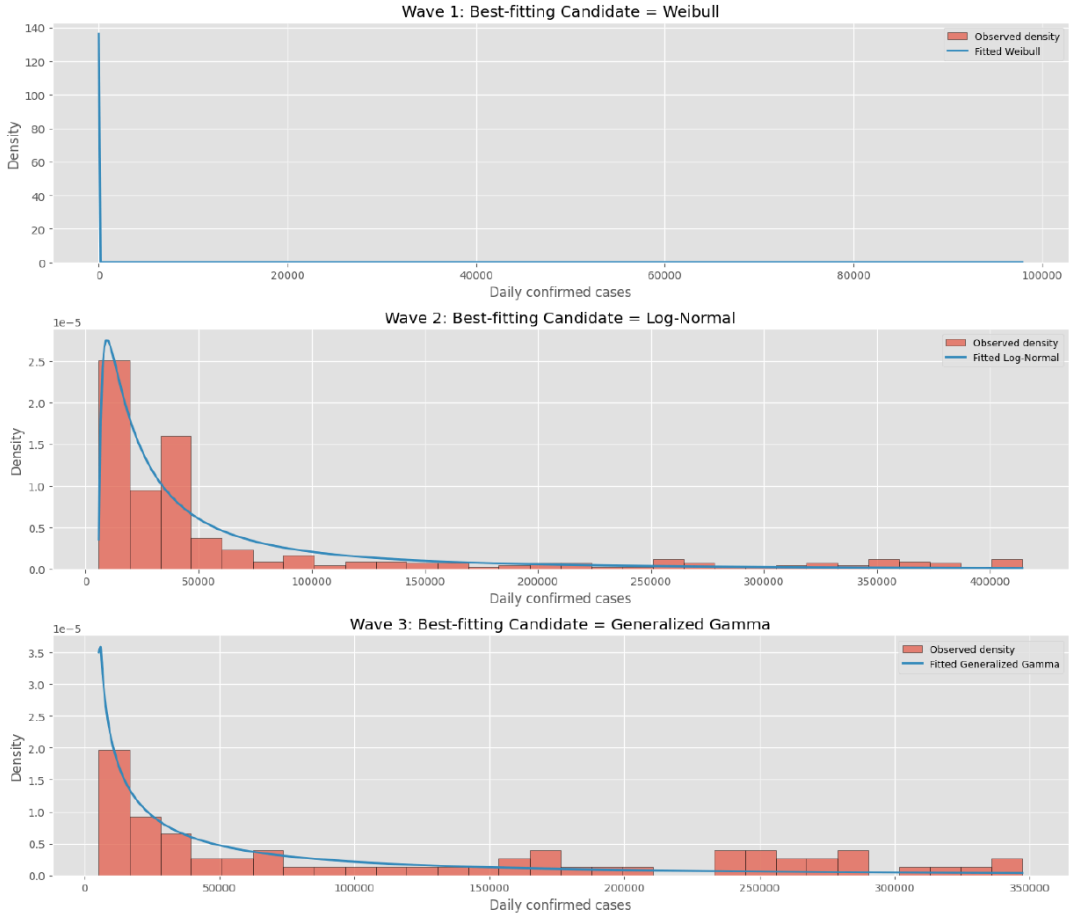

Figure 6: India wave-wise fitted candidate distributions for positive daily confirmed COVID-19 cases.

the United States and Japan, Inverse Gaussian distribution is selected for Brazil, Generalised Gamma distribution is selected for South Africa, and Log-Normal distribution is selected for the United Kingdom. This indicates that different countries have different distributional structures of daily cases. Therefore, one common distribution cannot be used for all countries. The Normal distribution performs poorly, which shows that a symmetric distribution is not suitable for COVID-19 daily case data. The Exponential distribution is also generally weak because it is too simple to represent the full epidemic case pattern.

The differences in selected distributions may be explained by differences in epidemic timing, wave structure, reporting intensity, and upper-tail behaviour across countries. Countries with long low-incidence periods followed by sharp high-incidence peaks may show stronger right-skewness and heavier upper tails. Countries with repeated moderate waves may show different dispersion and tail structures. Therefore, the selected distributions should not be interpreted only as mathematical results, but also as reflections of different empirical epidemic trajectories and reporting patterns. However, because the data are based on reported

confirmed cases, these differences may also reflect differences in testing intensity, case detection, and surveillance systems.

The results should be interpreted carefully because confirmed cases are not the same as total infections. Confirmed cases depend on testing availability, reporting system, case definition and surveillance capacity. Therefore, the observed differences between countries may be due to both the actual epidemic pattern and the reporting process. Still, the use of daily cases per million helps to reduce the effect of population size and gives a better basis for cross-country comparison. However, per-million scaling cannot remove the differences in testing and reporting practices. These limitations are consistent with earlier studies showing that cross-country COVID-19 comparisons may be affected by testing, surveillance, and reporting systems [6, 16, 19].

For India, the wave-wise analysis gives important results. Wave 1, Wave 2 and Wave 3 are not similar in magnitude and distributional shape. Wave 1 is best approximated by the Weibull distribution, Wave 2 by the Log-Normal distribution and Wave 3 by the Generalised Gamma distribution. This change in best-fitting distribution shows that India's COVID-19 case pattern changed from one wave to another. Hence, the complete Indian time series should not be treated as a single homogeneous distribution.

Wave 2 shows the strongest upper-tail behaviour and the highest maximum daily case count. This indicates that Wave 2 was not only larger in total burden, but also more extreme in distributional structure. Wave 3 has the highest mean and median daily cases, but it has a short duration. Therefore, the interpretation of Wave 3 should be done carefully because its distribution may be affected by the selected wave boundary and the small number of observations.

The bootstrap KS diagnostics also show that the selected distributions should be interpreted cautiously. In most datasets, the selected distributions provide useful empirical approximations. However, for the United Kingdom and Indian Wave 2, the bootstrap KS results indicate inadequate fit. This means that the selected Log-Normal distribution in these two cases is only the best relative candidate among the evaluated distributions, and should not be interpreted as an adequate parametric representation of the observed empirical distribution. Hence, the distributional results should be understood as comparative and descriptive, not as proof of exact parametric laws.

Overall, the findings support the use of flexible positive-support distributions for daily confirmed COVID-19 case data. The results also show that distribution pattern analysis gives useful information beyond total cases and average incidence. It helps to identify asymmetry, dispersion, extremity, and wave-specific behaviour in reported COVID-19 case series.

6. LIMITATIONS

This study has some limitations. First, confirmed COVID-19 cases do not represent the total number of infections. Confirmed cases depend on testing availability, reporting practice, surveillance capacity, and national case definitions. Therefore, the observed distributional patterns are patterns in reported confirmed cases, not complete SARS-CoV-2 infections [4, 6].

Second, cross-country comparison is affected by differences in health-care systems, reporting systems, testing policies, and public-health interventions. Per-million scaling adjusts for population size, but it cannot remove differences in testing and reporting practices. Therefore, the country-wise findings should be interpreted as differences in reported case distributions, not as direct differences in true infection distributions.

Third, zero observations are excluded from parametric distribution fitting because the evaluated candidate distributions are continuous positive-support distributions. This may introduce bias if zero-case days are epidemiologically meaningful. The issue is more relevant for Brazil and South Africa, where 5.05% of observations are excluded, and for Indian Wave 1, where 8.13% of observations are excluded. Therefore, the fitted distributions are conditional positive-case distributions. Zero-inflated or hurdle models may be useful in future studies to model zero and positive observations together.

Fourth, the fitted distributions are interpreted only as best-fitting candidates among the selected distributions. They are not interpreted as true epidemic-generating distributions. Daily COVID-19 case data are temporally dependent and may show serial correlation and temporal clustering. The present analysis does not model autocorrelation or transmission dynamics. Hence, the distributional models are used only as empirical marginal approximations of reported daily cases.

Fifth, the India wave-wise analysis depends on the selected wave boundaries. Wave 3 contains only 67 observations, and therefore, the fitted distribution for Wave 3 should be interpreted carefully. The small sample size may affect the stability of parameter

estimates and goodness-of-fit diagnostics. This concern is particularly important because the selected Wave 3 distribution is the Generalised Gamma distribution, which has greater parameter flexibility.

Finally, adequate distributional fit is not obtained for all datasets. In particular, the United Kingdom and Indian Wave 2 show inadequate fit under the bootstrap KS diagnostic. Therefore, the selected Log-Normal distribution in these cases should be interpreted only as the best relative candidate among the evaluated distributions, and not as an adequate distributional model.

7. CONCLUSION

In this study, we studied the distribution pattern of daily confirmed COVID-19 cases. The analysis is carried out for six countries: India, the United States, Brazil, the United Kingdom, South Africa and Japan. Along with this, an India-specific analysis is performed separately for three epidemic waves. The main aim was to understand how daily COVID-19 cases are distributed across countries and across Indian epidemic waves.

From the descriptive analysis, it is observed that daily COVID-19 case distributions are right-skewed in all selected countries. In all countries, the mean value is greater than the median value, which shows that high-incidence days affect the average case count. Skewness, kurtosis and robust statistics also indicate that the case distributions are not symmetric. Hence, the mean value alone is not sufficient to describe the distributional behaviour of daily cases.

The country-wise distribution fitting results show that no single distribution is best for all countries. Log-Logistic distribution is selected for India, the United States and Japan, Inverse Gaussian distribution is selected for Brazil, Generalised Gamma distribution is selected for South Africa, and Log-Normal distribution is selected for the United Kingdom under the composite ranking. However, the United Kingdom result should be interpreted cautiously because the bootstrap KS diagnostic indicates inadequate fit. The Normal distribution does not perform well, which shows that a symmetric distribution is not suitable for modelling daily COVID-19 case data. As a result, flexible positive-support and right-skewed distributions are more useful for describing the empirical distribution of positive daily cases, but selected distributions should be treated as relative empirical approximations.

For India, the wave-wise analysis shows that the three epidemic waves have different distributional behaviour. Wave 1 is best approximated by the Weibull distribution, Wave 2 by the Log-Normal distribution and Wave 3 by the Generalised Gamma distribution under the composite ranking. However, the Wave 2 Log-Normal fit should be interpreted only as the best relative candidate because the bootstrap KS diagnostic indicates inadequate fit. The Wave 3 result should also be interpreted carefully because it is based on only 67 observations and the selected Generalised Gamma distribution has greater parameter flexibility. These results indicate that the COVID-19 case pattern in India changed across epidemic phases. Hence, analysing the complete Indian series as a single distribution may hide important wave-specific patterns.

Overall, the study shows that COVID-19 daily confirmed cases are heterogeneous, asymmetric, and distributionally different across countries and Indian epidemic waves. Therefore, distribution pattern analysis provides useful statistical information beyond total cases and mean case counts.

REFERENCE

[1] Giménez V, Thieme C, Prior D, Tortosa-Ausina E. International comparisons of COVID-19 pandemic management: What can be learned from activity analysis techniques? Omega 2024; 122: 102954. https://doi.org/10.1016/j.omega.2023.102966

[2] Sun MW, Troxell D, Tibshirani R. Public health factors help explain cross-country heterogeneity in excess death during the COVID-19 pandemic. Scientific Reports 2023; 13: 16196. https://doi.org/10.1038/s41598-023-43407-0

[3] Guharay S, *et al*. A data-driven approach to study temporal characteristics of COVID-19 infection and death time series for twelve countries across six continents. BMC Medical Research Methodology 2025. https://doi.org/10.1186/s12874-024-02423-y

[4] Mathieu E, Ritchie H, Rodés-Guirao L, Appel C, Gavrilov D, Giattino C, Hasell J, Macdonald B, Dattani S, Beltekian D, Ortiz-Ospina E, Roser M. Coronavirus pandemic (COVID-19). Our World in Data 2020. https://ourworldindata.org/coronavirus

[5] Kumar G, Bhalla A, Mukherjee A, Turuk A, Talukdar A, Mukherjee S, *et al*. Post COVID sequelae among COVID-19 survivors: Insights from the Indian National Clinical Registry for COVID-19. BMJ Global Health 2023; 8: e012245.

[6] Alvarez E, Padgett D, Garnier R, Mattioli M, Engebretsen E. Limitations of COVID-19 testing and case data for evidence-informed health policy and practice. Health Research Policy and Systems 2023; 21: 11. https://doi.org/10.1186/s12961-023-00963-1

[7] Cavanaugh JE, Neath AA. Akaike Information Criterion, The: Background, derivation, properties, and refinements. In: Lovric M, ed. International Encyclopedia of Statistical Science 2025; 35-41. https://doi.org/10.1007/978-3-662-69359-9_685

[8] Charniga K, Park SW, Akhmetzhanov AR, Cori A, Dushoff J, Funk S, Gostic KM, Linton NM, Lison A, Overton CE, Pulliam JRC, Ward T, Cauchemez S, Abbott S. Best practices for estimating and reporting epidemiological delay distributions of infectious diseases. PLOS Computational Biology 2024; 20(10): e1012520. https://doi.org/10.1371/journal.pcbi.1012520

[9] Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Chapman & Hall/CRC 1993. https://doi.org/10.1007/978-1-4899-4541-9

[10] Liang CW, Lv QY, Chen ZG, Xu B, Lai YS, Zhang Z. Model-inferred timing and infectious period of the chickenpox

outbreak source. BMC Infectious Diseases 2024; 24(1): 1257. https://doi.org/10.1186/s12879-024-10127-3

[11] Massey FJ Jr. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association 1951; 46(253): 68-78. https://doi.org/10.1080/01621459.1951.10500769

[12] McAloon C, Collins Á, Hunt K, Barber A, Byrne AW, Butler F, Casey M, Griffin J, Lane E, McEvoy D, Wall P, Green M, O'Grady L, More SJ. Incubation period of COVID-19: A rapid systematic review and meta-analysis of observational research. BMJ Open 2020; 10(8): e039652. https://doi.org/10.1136/bmjopen-2020-039652

[13] National Institute of Standards and Technology. Kolmogorov-Smirnov goodness-of-fit test. NIST/SEMATECH e-Handbook of Statistical Methods 2012. https://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm

[14] Negi S, *et al*. Trend of viral load during the first, second, and third wave of COVID-19 in the Indian Himalayan region: An observational study of Uttarakhand state. Frontiers in Microbiology 2024; 14: 1279632. https://doi.org/10.3389/fmicb.2023.1279632

[15] Niveditha D, Khan S, Khilari A, Nadkarni S, Bhalerao U, Kadam P, *et al*. A tale of two waves: Delineating diverse genomic and transmission landscapes driving the COVID-19 pandemic in Pune, India. Journal of Infection and Public Health 2023; 16(9): 1417-1426. https://doi.org/10.1016/j.jiph.2023.06.004

[16] Padget M, Adam P, Rebolledo J, Riccardo F, Riess M, Rusu LC, Che D, Coignard B. A comparison of COVID-19 incidence rates across six European countries in 2021. Eurosurveillance 2023; 28(40): 2300088. https://doi.org/10.2807/1560-7917.ES.2023.28.40.2300088

[17] Portet S. A primer on model selection using the Akaike Information Criterion. Infectious Disease Modelling 2020; 5: 111-128. https://doi.org/10.1016/j.idm.2019.12.010

[18] Steinskog DJ, Tjøstheim DB, Kvamstø NG. A cautionary note on the use of the Kolmogorov-Smirnov test for normality. Monthly Weather Review 2007; 135(3): 1151-1157. https://doi.org/10.1175/MWR3326.1

[19] Tancredi S, Cullati S, Chiolero A. Surveillance bias in the assessment of the size of COVID-19 epidemic waves: A case study. Public Health 2024; 234: 98-104. https://doi.org/10.1016/j.puhe.2024.06.006

[20] Pawitan Y. In All Likelihood: Statistical Modelling and Inference Using Likelihood. Oxford University Press 2001. https://doi.org/10.1093/oso/9780198507659.001.0001

[21] Schwarz G. Estimating the dimension of a model. The Annals of Statistics 1978; 6(2): 461-464. https://doi.org/10.1214/aos/1176344136

Received on 07-05-2026 Accepted on 06-06-2026 Published on 01-07-2026

https://doi.org/10.6000/1929-6029.2026.15.25