УДК 519.23 : 616.98 ISSN 1995-5499
DOI: https://doi.org/10.17308/sait/1995-5499/2023/3/134-142
Поступила в редакцию 18.08.2023
Подписана в печать 30.09.2023

ПОСТРОЕНИЕ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ ПО ДАННЫМ СМИ О COVID-19

© 2023 Г. А. Курина[1,2]✉, М. Ю. Зиновьева[1], Е. А. Золотарева[1]

[1]Воронежский государственный университет
Университетская пл., 1, 394018 Воронеж, Российская Федерация
[2]Федеральный исследовательский центр «Информатика и управление» РАН
ул. Вавилова, 44/2, 119333 Москва, Российская Федерация

Аннотация. В данной статье выдвигается гипотеза о нормальном распределении количества пострадавших во время пандемии COVID-19, производится построение кривых нормального распределения для выделяемых в этот период так называемых волн. За основу берутся данные о пандемии, известные из средств массовой информации, а именно число заразившихся, умерших и выздоровевших. Все необходимые сведения находятся в открытом доступе. Для обработки информации, собранной по вышеуказанным показателям, используется эвристическая формула Стёрджесса, применяемая для определения «оптимального» числа интервалов разбиения области значений рассматриваемой случайной величины. Для полноты исследования анализируются данные по количеству заразившихся, умерших и выздоровевших, собранные для каждого из трёх показателей по разным странам с учетом временных рамок конкретных волн пандемии COVID-19. Для подтверждения гипотезы о нормальном распределении применяется критерий согласия Пирсона, также называемый критерием χ^2. В подавляющем большинстве исследованных случаев статистические данные не дают оснований отвергнуть гипотезу о нормальном распределении каждого из интересующих нас показателей в отдельности с учетом временных рамок соответствующих волн. В тексте статьи подробно описаны все шаги, необходимые для проверки справедливости выдвинутого предположения, приведены все используемые встроенные функции приложения Microsoft Office Excel, призванные оптимизировать процесс работы с большими объемами данных и визуализировать полученные результаты для большей наглядности. В качестве основного примера рассматривается статистика по Канаде, а именно по первой волне заболеваемости, пришедшейся на март — июнь 2020 года.

Ключевые слова: нормальное распределение, формула Стёрджесса, критерий согласия Пирсона, COVID-19.

ВВЕДЕНИЕ

Число работ, посвящённых методам прогнозирования инфекционной заболеваемости, в настоящее время быстро растёт благодаря появлению больших объёмов статистики, доступной для анализа. Во время эпидемии лицам, принимающим решения в области общественного здравоохранения, необходимы точные прогнозы числа случаев заболевания в будущем, чтобы контролировать распространение новых случаев и обеспечивать эффективное планирование ресурсов для удовлетворения потребностей и возможностей больниц. В [1] перечислены важные направления разработки математических моделей распространения заболевания: классические аналитические модели, детерминированные и стохастические, а также современные имитационные модели, сетевые и агентные.

В последние годы главным объектом исследований для учёных всего мира стал COVID-19 (см., например, [2–5]).

✉ Курина Галина Алексеевна
e-mail: kurina@math.vsu.ru

Заметную роль во многих областях науки играет так называемое нормальное распределение, плотность вероятности которого определяется функцией Гаусса

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2},$$

где a — математическое ожидание, σ — среднее квадратическое отклонение [6, 7]. Особое значение этого распределения в теории и практике основывается в значительной степени на центральной предельной теореме [8]. Интересное обсуждение нормального распределения приведено в [9].

Как указано в [10], один из основоположников медицинской статистики Уильям Фарр, по-видимому, был первым, кто установил, что во время эпидемии оспы в Англии и Уэльсе в 1837–1839 годах график числа смертей за квартал следовал примерно колоколообразной или нормальной кривой.

Объектом изучения в настоящей статье является собирательная статистика числа заразившихся, умерших и выздоровевших от COVID-19 по ряду стран. Данные для исследования страны * были получены с сайта

https://horoshotam.ru/*/coronavirus,

где * — русское название страны, записанное строчными латинскими буквами. Например, в случае Канады надо вместо * вставить kanada.

Цель данного исследования — проверить выдвинутую гипотезу о том, что распространение COVID-19 в пределах промежутка времени, соответствующего так называемой волне, в различных странах близко к нормальному распределению. Левый конец изучаемого промежутка соответствует дате, когда начинается существенное возрастание исследуемого показателя, а правый — дате, когда заканчивается заметное убывание этого показателя.

Для численного эксперимента использовалась программа Microsoft Excel.

1. СХЕМА ИССЛЕДОВАНИЯ

Сначала находим данные для выбранной страны с указанного во Введении сайта. Формат данных следующий — каждой дате соответствует число случаев по выбранному показателю. Затем выбираем исследуемый временной интервал, соответствующий одной волне.

Далее по данным выборки вычисляем оценки параметров нормального распределения, а именно, выборочное среднее $\overline{x}_B$ и исправленное выборочное среднее квадратическое отклонение s, используя встроенные в Excel функции «*СРЗНАЧ*()» и «*СТАНДОТКЛ*()» соответственно.

1.1. Теоретические частоты

Для проверки гипотезы о предполагаемом нормальном законе распределения промежуток изменения исследуемого показателя разбиваем на k интервалов с концами x_{1i}, x_{2i}, $i=\overline{1,k}$. Обозначим через n_i количество значений анализируемого показателя в i-м полуинтервале $[x_{1i},x_{2i})$, при $i=k$ вместо полуинтервала рассматриваем замкнутый интервал. Числа n_i называются эмпирическими частотами, $n=\sum_{i=1}^{k}n_i$ — объемом выборки.

Выравнивающие (теоретические) частоты n_i' будем вычислять по формуле

$$n_i'=n(\Phi(t_{2i})-\Phi(t_{1i})), \qquad (1)$$

где

$$t_{ji}=\frac{x_{ji}-\overline{x}_B}{s},\quad j=1,2,$$

а значения функции Лапласа

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{x^2}{2}}dx$$

находятся по таблице (в Excel можно использовать встроенную функцию *НОРМСТРАСП*()). Найденные значения n_i' будем округлять до двух знаков после запятой.

Обозначим через x_i середину интервала (x_{1i},x_{2i}). По точкам (x_i,n_i') строится кривая нормального распределения. Точки с координатами (x_i,n_i) далее будут изображаться на графиках темными точками.

1.2. Критерий согласия Пирсона

В этой статье для проверки гипотезы о предполагаемом законе распределения используется критерий согласия Пирсона, называемый также критерием χ^2 (см., например, [6, 7]).

Для этого вычисляется наблюдаемое значение критерия:

$$\chi^2_{набл} = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i}, \qquad (2)$$

где n_i — эмпирические, а n'_i — теоретические частоты.

Затем по заданному уровню значимости α и числу степеней свободы $m = k - 1 - r$, где k — число интервалов, на которые разбита область значений исследуемой случайной величины, r — число неизвестных параметров предполагаемого закона распределения, которые оцениваются по данным выборки, находится критическая точка $\chi^2_{кр}$ («*ХИ2ОБР*()»).

Если $\chi^2_{набл} \le \chi^2_{кр}$, то нет оснований отвергнуть гипотезу о предполагаемом распределении, если $\chi^2_{набл} > \chi^2_{кр}$, то гипотеза отвергается.

В нашем случае проверки гипотезы о нормальном распределении имеются два параметра (математическое ожидание и среднее квадратическое отклонение), поэтому $r = 2$ и число степеней свободы $m = k - 3$. В последующих расчётах будем использовать значение $\alpha = 0{,}05$.

Статистические свойства критериев зависят как от того, каким образом область определения случайной величины разбивается на интервалы, так и от выбора числа интервалов группирования. Рекомендуемое в различных источниках количество интервалов группирования, используемое, например, при проверке статистических гипотез с помощью критерия согласия Пирсона, колеблется в очень широких пределах. Большинство из рекомендуемых формул для оценки числа интервалов носит эмпирический характер. Естественно, что определение количества интервалов связывается с объемом выборки. В [11] на основании разных источников приводится обзор различных методов выбора числа интервалов k.

Приведем некоторые советы из [11]. При выборе интервалов равной длины наиболее часто рекомендуется, чтобы количество наблюдений, попавших в интервал, было не менее 10. На практике допустимо, чтобы количество наблюдений в крайних интервалах было менее 5. В случае использования критерия согласия Пирсона допускается уменьшение ожидаемых частот попадания наблюдений для одного или двух интервалов до 1 и даже ниже.

Во многих источниках можно найти упоминание эвристической формулы Стёрджесса для определения «оптимального» числа интервалов [12]

$$k = 1 + \lceil \log_2 n \rceil,$$

где $[x]$ означает целую часть числа x.

Перечисленные выше рекомендации будем использовать при выборе числа интервалов.

2. РЕЗУЛЬТАТЫ ЧИСЛЕННОГО ЭКСПЕРИМЕНТА

В этом разделе будут представлены результаты анализа данных СМИ по трем показателям для первой волны заболеваемости в Канаде.

2.1. Построение кривой нормального распределения

Немаловажную роль при построении кривой нормального распределения по эмпирическим данным имеет выбор временных границ волны, а также числа разбиений. В данной работе число разбиений определяется по формуле Стёрджесса, и интервалы с малым числом эмпирических частот объединяются.

Следует отметить, что при анализе данных по выздоравливаемости и смертности рамки временного интервала для волны сдвигаются вперед относительно выбранных дат для анализа зараженности. Данные показатели являются зависимыми от показателя зараженности.

Рассмотрим построение кривой нормального распределения по данным с указанного во Введения сайта, касающимся числа заразившихся коронавирусом в Канаде в первую волну.

Таблица 1. Данные по числу заразившихся в Канаде за период 22.03.20–05.06.20 [Table 1. Data on the number of people infected in Canada for the period 03/22/20–06/05/20]

Даты	Данные	Даты	Данные
22.03.	192	29.04.	1571
23.03.	619	30.04.	1639
24.03.	702	01.05.	1825
25.03.	461	02.05.	1653
26.03.	791	03.05.	2760
27.03.	640	04.05.	1298
28.03.	894	05.05.	1274
29.03.	744	06.05.	1450
30.03.	1128	07.05.	1426
31.03.	1164	08.05.	1512
01.04.	1119	09.05.	1268
02.04.	1552	10.05.	1146
03.04.	1092	11.05.	1133
04.04.	1537	12.05.	1176
05.04.	1600	13.05.	1121
06.04.	1155	14.05.	1123
07.04.	1230	15.05.	1212
08.04.	1541	16.05.	1251
09.04.	1327	17.05.	1138
10.04.	1383	18.05.	1070
11.04.	1170	19.05.	1040
12.04.	1065	20.05.	1030
13.04.	1297	21.05.	1182
14.04.	1383	22.05.	1156
15.04.	1316	23.05.	1141
16.04.	1727	24.05.	1078
17.04.	1821	25.05.	1012
18.04.	1456	26.05.	936
19.04.	1673	27.05.	872
20.04.	1773	28.05.	993
21.04.	1593	29.05.	906
22.04.	1768	30.05.	772
23.04.	1920	31.05.	757
24.04.	1778	01.06.	758
25.04.	1466	02.06.	705
26.04.	1541	03.06.	675
27.04.	1605	04.06.	641
28.04.	1526	05.06.	609

Используя функции «*СРЗНАЧ*()», «*СТАНДОТКЛ*()», получаем следующие значения: $\overline{x}_B = 1224{,}45$, $s = 407{,}06$ (в результатах вычислений будем оставлять два знака после запятой).

Для определения диапазона изменения значений случайной величины найдём максимальное и минимальное значение числа заболевших, используя встроенные в Excel функции «*MAX*()» и «*МИН*()». В результате получаем $x_{\max} = 2706$, $x_{\min} = 192$. Диапазон [192; 2706].

Учитывая формулу Стёрджесса при $n = 76$, возьмем число интервалов разбиения одинаковой длины равное 7.

Теперь найдём эмпирические частоты n_i для каждого полуинтервала $[x_{1i}, x_{2i})$, то есть, считаем число заразившихся, попадающих в этот полуинтервал. Для последнего промежутка разбиения рассматриваем замкнутый интервал. Поскольку эмпирические частоты в шестом и седьмом интервалах равны соответственно 0 и 1, то объединяем два последних интервала в один.

Результаты представлены в следующей таблице, где в 1-м столбце указаны номера интервалов разбиения.

Таблица 2. Эмпирические и теоретические частоты для данных о заразившихся [Table 2. Empirical and theoretical frequencies for data on infected people]

№	x_{1i}	n_i	t_{1i}	n_i'
1	192	2	–2,54	3,45
2	558,86	15	–1,64	13,72
3	925,71	28	–0,73	25.46
4	1292,57	22	0,17	22.11
5	1659,43	8	1,07	8.98
6	2026,29	1	1,97	1,85

Заметим, что $x_{2i} = x_{1(i+1)}$, $i = \overline{1,5}$, $x_{26} = x_{\max} = 2706$. Аналогичные формулы имеют место для t_{2i}. В частности, $t_{26} = 3{,}77$.

В предпоследнем столбце последней таблицы представлены значения

$$t_{1i} = \frac{x_{1i} - \overline{x}_B}{s},$$

а в последнем столбце — теоретические частоты, вычисленные по формуле (1) с точностью до двух знаков после запятой с использованием «*НОРМСТРАСП*()».

На приведенных в статье графиках по горизонтальной оси указываются номера интервалов разбиения области значений исследуемого показателя, а по вертикальной — соответствующие частоты.

Используя полученные данные, строим кривую нормального распределения, соединяя точки с координатами (x_i, n_i').

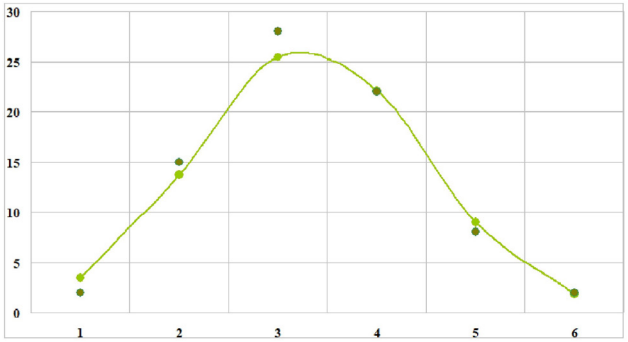

Рис. 1. Кривая нормального распределения, построенная по данным о числе заразившихся в Канаде в первую волну (22.03.20–05.06.20) [Fig. 1. A normal distribution curve constructed using data on number of people infected in Canada during the first wave (03/22/20–06/05/20)]

Как уже говорилось, темными точками на рисунках обозначаются точки с координатами (x_i, n_i). Отметим, что здесь и на остальных рисунках эти точки не отражают всю информацию о выборке, так как не учитываются конкретные значения всех вариант из рассматриваемого частичного интервала и их частоты.

Подобным образом исследовались данные по числу умерших и выздоровевших в первую волну в Канаде соответственно за промежуток времени 06.04.20–14.06.20 и 03.04.20–28.06.20. Таблицы данных аналогичные табл. 1 здесь не приводятся, поскольку эти данные можно найти на указанном во Введении сайте. Приведем только соответствующие аналоги табл. 2 и графики кривых нормального распределения, построенных по эмпирическим данным.

Используя данные о числе умерших за промежуток времени 06.04.20–14.06.20, имеем диапазон числа умерших в этом промежутке $[27, 222]$. $\overline{x}_B = 112{,}37$, $s = 48{,}9$. По формуле Стёрджесса при $n = 70$ получаем число интервалов разбиения $k = 7$.

Таблица 3. Эмпирические и теоретические частоты для данных об умерших [Table 3. Empirical and theoretical frequencies for death data]

№	x_{1i}	n_i	t_{1i}	n_i'
1	27	7	–1,75	5,55
2	54,86	15	–1,18	10,66
3	82,71	12	–0,61	14,93
4	110,57	14	–0,04	15,23
5	138,43	9	0,53	11,34
6	166,29	11	1,10	6,15
7	194,14	2	1,67	2,43

Здесь $x_{2i} = x_{1(i+1)}$, $i = \overline{1,6}$, $x_{27} = 222$. Аналогичные формулы имеют место для t_{2i}. В частности, $t_{27} = 2{,}24$.

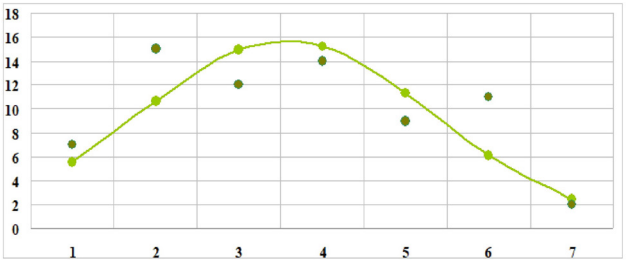

Рис. 2. Кривая нормального распределения, построенная по данным о числе умерших в Канаде в первую волну (06.04.20–14.06.20) [Fig. 2. A normal distribution curve constructed using data on number of deaths in Canada during the first wave (04/06/20–06/14/20)]

Используя данные о выздоровевших за промежуток времени 03.04.20–28.06.20, имеем диапазон числа выздоровевших в этом промежутке $[207{,}1434]$, $\overline{x}_B = 738{,}07$, $s = 250{,}53$. По формуле Стёрджесса при $n = 87$ получаем число интервалов разбиения $k = 7$. Поскольку эмпирическая частота в седьмом интервале равна 2, то объединяем два последних интервала в один.

Таблица 4. Эмпирические и теоретические частоты для данных о выздоровевших [Table 4. Empirical and theoretical frequencies for recovered data]

№	x_{1i}	n_i	t_{1i}	n'_i
1	207	7	–2,12	5,29
2	382,29	12	–1,42	13,73
3	557,57	23	–0,72	22,28
4	732,86	24	–0,02	22,59
5	908,14	13	0,68	14,32
6	1083,43	8	1,38	7,07

Здесь $x_{2i} = x_{1(i+1)}$, $i = \overline{1,5}$, $x_{26} = 1434$. Аналогичные формулы имеют место для t_{2i}. В частности, $t_{26} = 2,78$.

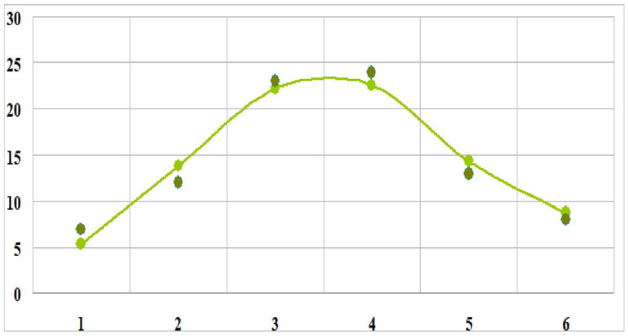

Рис. 3. Кривая нормального распределения, построенная по данным о числе выздоровевших в Канаде в первую волну (03.04.20–28.06.20) [Fig. 3. A normal distribution curve constructed using data on number of recoveries in Canada in the first wave (04/03/20–06/28/20)]

2.2. Проверка гипотезы о нормальном распределении

Проанализируем полученные результаты с помощью критерия согласия Пирсона.

Для первой волны заболевших в Канаде по формуле (2) при $k = 6$, используя «*СУММ*()», находим $\chi^2_{набл} = 1,10$. При уровне значимости $\alpha = 0,05$ и числе степеней свободы $m = 6 - 3 = 3$. используя «*ХИ2ОБР*()», находим критическую точку $\chi^2_{кр} = 7,81$.

Принимая во внимание полученные результаты для первой волны умерших в Канаде, вычисляем $\chi^2_{набл} = 7,20$. При уровне значимости $\alpha = 0,05$ и числе степеней свободы $m = 7 - 3 = 4$ имеем $\chi^2_{кр} = 9,49$.

Используя полученные результаты для числа выздоровевших в Канаде в первую волну, находим $\chi^2_{набл} = 1,06$. Здесь $\chi^2_{кр} = 7,81$.

Поскольку во всех трех исследованных случаях $\chi^2_{набл} < \chi^2_{кр}$, то в силу критерия согласия Пирсона в этих случаях нет оснований отвергнуть гипотезу о нормальном распределении.

ЗАКЛЮЧЕНИЕ

В статье приведены кривые нормального распределения, построенные по эмпирическим данным для первой волны трех показателей (зараженность, смертность и выздоравливаемость) в Канаде. При этом использовалась эвристическая формула Стёрджесса для определения «оптимального» числа интервалов разбиения области значений для изучаемого показателя. Для этих случаев представлены также результаты исследования при помощи критерия согласия Пирсона вопроса о правильности предположения о нормальном распределении. Для всех трех рассмотренных случаев нет оснований отклонить гипотезу о нормальном распределении.

Отметим интересную, на наш взгляд, закономерность. У. Фарр указал закон изменения числа смертей от оспы для показателей за один квартал. Выбранная нами интуитивно из косвенных соображений продолжительность волн охватывает промежуток около трех месяцев.

Кроме Канады, были исследованы данные по трём рассматриваемым показателям для нескольких волн в Болгарии, Бразилии, Великобритании, Германии, Доминиканской республике, Индии, Италии, Кении, Китае, Норвегии, России, США, Турции, Франции, Чили, Швейцарии. В подавляющем большинстве исследованных случаев при уровне значимост $\alpha = 0,05$ в силу критерия согласия Пирсона нет оснований отвергнуть гипотезу о нормальном распределении отдельного показателя.

Одни из основных факторов, которые могут помешать подтвердить гипотезу о нормальном распределении: «пики» и «плато» показателей. «Пик» показателя — день, в который произошел всплеск числа случаев, а затем — резкий

спад. Это может быть вызвано спецификой сбора данных о числе случаев медучреждениями и лабораториями. «Плато» показателя — период времени, в который большое число случаев находится в фиксированном промежутке значений этого показателя.

Кривые нормального распределения, построенные по эмпирическим данным о COVID-19 для некоторых стран, приведены в магистерских диссертациях Максимова Ю. М. (2021), Никулина Р.А. (2022) и Лебедевой И. С. (2023), выполненных на математическом факультете Воронежского государственного университета. Никулиным Р. А. была разработана программа на языке C#, использующая следующие задаваемые входные данные: страна, анализируемый показатель (зараженность, смертность, выздоравливаемость), исследуемый временной промежуток (волна), число разбиений промежутка. В отличие от настоящей статьи, в этих работах для промежутка значений анализируемого показателя рассматривалось различное число интервалов разбиения одинаковой длины, не связанное с формулой Стёрджесса.

БЛАГОДАРНОСТИ

Работа Г. А. Куриной поддержана РНФ, грант 21-11-00202.

Авторы выражают благодарность слушателям курса «Теория прогнозирования» за проведение расчётов для различных стран. Они также искренне признательны уважаемому Рецензенту за полезные замечания.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

СПИСОК ЛИТЕРАТУРЫ

1. *Кондратьев М. А.* Методы прогнозирования и модели распространения заболеваний / М. А. Кондратьев // Компьютерные исследования и моделирование. – 2013. – Т. 5, № 5. – С. 863–882.

2. *Романов Б. К.* Коронавирусная инфекция Covid – 2019 / Б. К. Романов // Безопасность и риск фармакотерапии. – 2020. – Т. 8, № 1. – С. 3–8. https://doi.org/10.30895/2312-7821-2020-8-1-3-8

3. Новая коронавирусная инфекция (COVID-19): клинико-эпидемиологические аспекты / В. В. Никифоров [и др.] // Архивъ внутренней медицины. – 2020. – Т. 10, № 2. – С. 87–93. https://doi.org/10.20514/2226-6704-2020-10-2-87-93

4. *Alzubadi H.* Modeling the Infection Disease (Covid-19) and the Effect of Vaccination // Applied Mathematics. – 2023. – Vol. 14, No 7. Available at: http://scrip.org/journal/paperinformaton.aspx?paperind=126271. – doi:10.4236/am.2023.147027

5. A New Mathematical Modeling of the Covid-19 Pandemic Including the Vaccination Campaign / M. Yavuz [et al.] // Open Journal of Modeling and Simulation. – 2021. – Vol. 9, No 3. Available at: http://scrip.org/journal/paperinformaton.aspx?paperind=110660. – doi: 10.4236/ojmsi.2021.93020

6. *Гмурман В. Е.* Теория вероятностей и математическая статистика / В. Е. Гмурман. – Москва : Высшая школа, 1999. – 479 с.

7. Теория вероятностей и математическая статистика / В. С. Мхитарян [и др.] – Москва: Московский международный институт эконометрики, информатики, финансов и права, 2003. – 130 с. https://www.studmed.ru/view/mhitaryan-vs-troshin-li-adamova-ev-shevchenko-kk-bambaeva-nya-teoriya-veroyatnostey-i-matematicheskaya-statistika_ee388028710.html

8. *Сенатов В. В.* Центральная предельная теорема: Точность аппроксимации и асимптотические разложения / В. В. Сенатов. – URSS, 2018. – 352 с. https://urss.ru/cgi-bin/db.pl?lang=Ru&blang=ru&page=Book&id=229391

9. *Талеб Н. Н.* Черный лебедь. Под знаком непредсказуемости / Н. Н. Талеб. – Москва : КоЛибри, Азбука-Аттикус, 2022. – 736 с.

10. *Martin W. J.* The Epidemic Curve of Smallpox // The Journal of Hygiene. – 1934. – Vol. 34, No 1. – P. 10–29.

11. *Лемешко Б. Ю., Чимитова Е. В.* О выборе числа интервалов в критериях согласия типа χ^2 / Б. Ю. Лемешко, Е. В Чимитова // Заводская ла-

боратория. Диагностика материалов. – 2003. – Т. 69. – С. 61–67. https://www.researchgate.net/publication/315333672_O_vybore_cisla_intervalov_v_kriteriah_soglasia_tipa_X2

12. *Sturgess H. A.* The choice of classic intervals / H. A. Sturgess // Journal of the American Statistical Association. – 1926. – Vol. 21, No 153. – P. 65–66.

Курина Галина Алексеевна — д-р физ.-мат. наук, проф., профессор кафедры математического анализа Воронежского государственного университета, Федеральный исследовательский центр «Информатика и управление» РАН. E-mail: kurina@math.vsu.ru
ORCID iD: https://orcid.org/0000-0002-1586-9943

Зиновьева Мария Юрьевна — магистрант 1-го года обучения кафедры математического анализа Воронежского государственного университета. E-mail: shkondamari@mail.ru
ORCID iD: https://orcid.org/0009-0000-4762-3772

Золотарева Евгения Андреевна — магистрант 2-го года обучения кафедры математического анализа Воронежского государственного университета. E-mail: zolotareva.1@mail.ru
ORCID iD: https://orcid.org/0009-0006-7835-2251

DOI: https://doi.org/10.17308/sait/1995-5499/2023/3/134-142 ISSN 1995-5499
Received 18.08.2023
Accepted 30.09.2023

CONSTRUCTING NORMAL DISTRIBUTION ACCORDING TO MEDIA DATA ABOUT COVID-19

© 2023 G. A. Kurina[1,2]✉, M. Yu. Zinov'eva[1], E. A. Zolotareva[1]

[1]Voronezh State University
1, Universitetskaya Square, 394018 Voronezh, Russian Federation
[2]Federal Research Center "Computer Science and Control" of Russian Academy of Science
44/2, Vavilova Street, 119333 Moscow, Russian Federation

Annotation. This paper deals with a hypothesis about the normal distribution of the quantity of victims during the COVID-19 pandemic, and normal distribution curves for the so-called waves identified during this period are constructed. The data on COVID-19 known from the media are taken as a basis, namely the quantity of infected, died and recovered. All necessary information is publicly available. To study the collected information regarding the above indicators, the Sturgess heuristic formula is used to determine the "optimal" number of intervals for partitioning the range of values of the random variable under consideration. For completeness, the study analyzes data on the quantity of infected, died and recovered, collected for each of the three indicators for different countries, taking into account the time frame of specific waves of the pandemic. To confirm the hypothesis of normal distribution, the Pearson goodness-of-fit test, also called the χ^2 test, is used. In the overwhelming majority of the cases studied, statistical data do not provide grounds to reject the hypothesis about the normal distribution of each of the indicators we are interested in separately, taking into account the time frame of the corresponding waves. The text of the paper shows in detail all steps necessary to verify the validity of the assumption made. All used built-in functions of the Microsoft Office Excel application are indicated, designed to optimize the process of working with large volumes of data and visualize the results for greater clarity. As a main example, statistics for Canada are considered, namely for the first wave of morbidity, which occurred in March – June 2020.

Keywords: normal distribution, Sturgess formula, Pearson's goodness-of-fit test, COVID-19.

✉ Курина Галина Алексеевна
e-mail: kurina@math.vsu.ru

CONFLICT OF INTEREST

The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.

REFERENCES

1. *Kondrat'ev M. A.* (2013) Metody prognozirovaniya i modeli rasprostraneniya zabolevanii [Methods of Forecasting and Models of the Spread of Diseases]. *Kompyuternye issledovaniya i modelirovanie*. 5 (5). P. 863–882. (in Russian)

2. *Romanov B. K.* (2020) Koronavirusnaya infekciya Covid-2019 [Coronavirus Disease Covid-2019]. *Bezopasnost' i risk farmakoterapii*. 8 (1). P. 3–8. (in Russian)

3. *Nikiforov V. V., Suranova T. G., Chernobrovkina T. Ya., Yankovskaya Y. D. and Burova S. V.* (2020) Novaya koronavirusnaya infekciya (COVID-19): kliniko-epidemiologicheskie aspekty [New Coronavirus Infection (COVID-19): Clinical and Epidemiological Aspects]. Arhiv vnutrennei mediciny. 10 (2). P. 87–93. (in Russian).

4. *Alzubadi H.* (2023) Modeling the Infection Disease (Covid-19) and the Effect of Vaccination. Applied Mathematics. 14 (7). https:/www.scrip.org/journal/paperinforcitaton.aspx?paperind=126271. – doi:10.4236/am.2023.147027

5. *Yavuz M., Coşar F., Günay F. and Özdemir F.* (2021) A New Mathematical Modeling of the Covid-19 Pandemic Including the Vaccination Campaign. *Open Journal of Modeling and Simulation*. 9 (3). https://www.scrip.org/journal/paperinformaton.aspx?paperind=110660. – doi: 10.4236/ojmsi.2021.93020

6. *Gmurman V. E.* (1999) Teoriya veroyatnostei i matematicheskaya statistika [Probability Theory and Mathematical Statistiks]. Moscow : Higher School. (in Russian)

7. *Mkhitaryan V. S., Troshin L. I., Adamova E. V., Shevchenko K. K. and Bambaeva N. Ya.* (2003) Teoriya veroyatnostei i matematicheskaya statistika [Probability Theory and Mathematical Statistics]. Moscow : Moskovskii mezhdunarodnyi institut ekonometriki, informatiki, finansov i prava publ. (in Russian)

8. *Senatov V. V.* (2018) Central'naya predel'naya teorema: Tochnost' approkcimacii i asymptoticheskie razlozheniya [Central Limit Theorem: Accuracy of Approximation and Asymptotic Expansions]. URSS. (in Russian)

9. *Taleb N. N.* (2022) Black Swan. Under the sign of unpredictability. Moscow, Hummingbird, ABC–Atticus.

10. *Martin W. J.* (1934) The Epidemic Curve of Smallpox. *The Journal of Hygiene*. 34 (1). P. 10–29.

11. *Lemeshko B. Yu. and Chimitova E. V.* (2003) O vybore chisla intervalov v kriteriyah soglasiya tipa χ^2 [About Choice of Intervals in Goodness-of-fit Test of χ^2 Type]. *Zavodskaya laboratoriya. Diagnostika materialov*. 69. P. 61–67. (in Russian)

12. *Sturgess H. A.* (1926) The Choice of Class-Interval. *Journal of the American Statistical Association*. 21 (153). P. 65–66.

Kurina Galina A. — dr. Phys.-Math. Sciences, professor, professor of department of mathematical analysis, Voronezh State University, Federal Research Center "Computer Science and Control" of Russian Academy of Science.
E-mail: kurina@math.vsu.ru
ORCID iD: https://orcid.org/0000-0002-1586-9943

Zinov'eva Mariya Yu. — 1st year master's student of department of mathematical analysis, Voronezh State University.
E-mail: shkondamari@mail.ru
ORCID iD: https://orcid.org/0009-0000-4762-3772

Zolotareva Jane A. — 2nd year master's student of department of mathematical analysis, Voronezh State University.
E-mail: zolotareva.1@mail.ru
ORCID iD: https://orcid.org/0009-0006-7835-2251