Estimation of the probability of daily fluctuations of incidence of COVID-19 according to official data (#52610)

First submission

Guidance from your Editor

Please submit by **18 Oct 2020** for the benefit of the authors (and your $200 publishing discount) .

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Author notes
Have you read the author notes on the guidance page?

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files
Download and review all files from the materials page.

3 Figure file(s)
2 Table file(s)

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- Impact and novelty not assessed. Negative/inconclusive results accepted. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Speculation is welcome, but should be identified as such.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

Estimation of the probability of daily fluctuations of incidence of COVID-19 according to official data

Andrey Gerasimov [Corresp., 1], **Elena Galkina** [1], **Elena Danilova** [1], **Irina Ikonnikova** [1], **Tamara Novoselova** [1], **Yuriy L Orlov** [2], **Irina Senenycheva** [1]

[1] Department of Medical Informatics and Statistics, I.M. Sechenov First Moscow State Medical University, Moscow, Russia

[2] Institute of Digital Medicine, I.M. Sechenov First Moscow State Medical University, Moscow, Russia

Corresponding Author: Andrey Gerasimov
Email address: andr-gerasim@yandex.ru

In addition to the overall morbidity and mortality rate, as well as the trends of their changes, information interesting for epidemiologists can be obtained from the analysis of the differences in the number of morbidity and deaths in adjacent days. Increased differences can be a result of both trends in morbidity or mortality changes and changes in diagnosis criteria and of the clustering of cases into groups, such as family hotbeds of disease or microepidemics in closed communities. When approbation of this technique on the data of COVID-19 for Poland, Moscow and Russia excluding Moscow an interesting phenomenon was found: for Moscow differences in the numbers of sick and dead in neighboring days statistically is significantly less than expected in the assumption of the independence and persistence of morbidity or mortality. Consequently, Moscow is adjusting the actual data to show a picture of stable morbidity.

Estimation of the probability of daily fluctuations of incidence of COVID-19 according to official data

Andrey Gerasimov[1], Elena Galkina[1], Elena Danilova[1], Irina Ikonnikova[1], Tamara Novoselova[1], Yuriy Orlov[1], Irina Senenycheva[1]

[1] I.M. Sechenov First Moscow State Medical University (Sechenov University), 2-4 Bolshaya Pirogovskaya Street, 119991 Moscow, Russia

* Correspondence author

e-mail: andr-gerasim@yansex.ru

Andrey Gerasimov 105043 Izmailovsky boulevard, 32/23, building 1, apt 78 +7(905)550-5084

Keywords

COVID-19, random fluctuations of incidence, mathematical methods of morbidity analysis

Abstract

In addition to the overall morbidity and mortality rate, as well as the trends of their changes, information interesting for epidemiologists can be obtained from the analysis of the differences in the number of morbidity and deaths in adjacent days. Increased differences can be a result of both trends in morbidity or mortality changes and changes in diagnosis criteria and of the clustering of cases into groups, such as family hotbeds of disease or microepidemics in closed communities. When approbation of this technique on the data of COVID-19 for Poland, Moscow and Russia excluding Moscow an interesting phenomenon was found: for Moscow differences in the numbers of sick and dead in neighboring days statistically is significantly less than expected in the assumption of the independence and persistence of morbidity or mortality. Consequently, Moscow is adjusting the actual data to show a picture of stable morbidity.

Introduction

An integral part of the analysis of actual morbidity data is to assess the correctness and comparability of official reporting data (Cooper et al., 2009; Isanaka S et al., 2016)

Differences in incidence may be related not only and not so much to differences in disease risk, but to differences in case detection and recording. Therefore, if in the first region the incidence is higher than in the second under similar conditions, it does not mean that doctors perform worse in the first region, the situation is often the opposite one.

Detecting a case and making a diagnosis is not something unambiguous.

First, for almost all infectious diseases, the proportion of manifest cases is low. Most cases are asymptomatic or unclear symptomatic, and for the most part remain undetected. Some cases are not detected by referrability, but by active detection efforts in foci and among at-risk groups (Abbott et al., 2017; Leung, Trapman & Britton, 2018).

Second, incidence depends on diagnosis criteria. For example, an infectious disease can be diagnosed both on symptoms and in case of presence of laboratory confirmation. In the latter case, a large part of them goes recorded as “acute respiratory disease” and “intestinal infection of unclarified epidemiology”.

The criteria for diagnosis can vary over time and between countries. For example, the decrease in the incidence of tuberculosis in Russia at a rate of about 10% observed in the last decade is a consequence of the constant change in diagnosis criteria with their fitting to WHO criteria, where only cases with active bacteriodisposition are considered as tuberculosis, while maintaining constant criteria for diagnosing, the incidence of tuberculosis in Russia would continue to increase (Yu WY et al., 2019; Герасимов 2018).

Fluctuations in morbidity can also be associated with organisational aspects. For example, in Brazil, the number of COVID-19 cases detected on Saturday and Sunday is about one and a half times lower than on weekdays. In Russia, there are no differences in the incidence of COVID-19 on weekdays and Sundays, but there is a three times difference in the number of recovered due to the fact that on Saturday and Sunday there are no dismisses from hospitals.

Unfortunately, among the factors influencing official morbidity, there is also a desire to show a picture better than it is in reality, hide flaws and errors and “report nicely”. This is usually found in a decrease in morbidity and mortality. An example is the legendary statement of the Belarusian leader that “we have no deaths from COVID, we have deaths with COVID”.

The epidemic process is a random process, therefore, the dynamics of the incidence includes random fluctuations (Black et al., 2009; Krause et al., 2018; Simões, Telo da Gama & Nunes 2008)

However, an analysis of official COVID-19 incidence data revealed another phenomenon: an overly stable incidence, in which the number of cases detected per day hardly changes.

Materials and methods

Let A be the number of cases detected over a period of time, including per day. Then if the cases are independent and the number of cases is low compared to the overall population, then A is distributed by Poisson (Герасимов 2014). For the Poisson distribution, variance equals mathematical expectation. Therefore, if x_1, x_2 are two independent observations of the Poisson distribution with the average λ, then $E(x_1 - x_2) = E(x_1) - E(x_2) = \lambda - \lambda = 0$ and $D(x_1 - x_2) = D(x_1) + D(x_2) = \lambda + \lambda = 2\lambda$.

Besides, for a sufficiently large mathematical expectation, the Poisson distribution is close to the normal distribution (Stuart 2009). Therefore, when increasing the mathematical expectation of the Poisson distribution, the distribution of value $(x_1 - x_2)$ tends to a normal

distribution with a mathematical expectation equal to zero and variance, equal to 2λ. Consequently, when increasing λ, the distribution approximates distribution $2\lambda\chi_1^2$,

However, the mathematical expectation of the number of sick is unknown to us. If x_1, x_2 are two independent observations of the Poisson distribution with mathematical expectation λ, then $\Delta = \frac{(x_1 - x_2)^2}{x_1 + x_2}$ is not distributed as χ_1^2, since, first, the Poisson distribution is not exactly the same as normal and secondly, in expression $\frac{(x_1 - x_2)^2}{x_1 + x_2}$ the same values x_1, x_2 are used both to estimate variance, and to estimate mathematical expectation.

The value distribution function $\Delta = \frac{(x_1 - x_2)^2}{x_1 + x_2}$ can be calculated as

$$F_\Delta(x) = \sum_{k\geq 0, n\geq 0, k+n>0, \Delta\leq x} \frac{e^{-k}}{k!} e^{-\lambda} \frac{e^{-n}}{n!} e^{-\lambda} = \sum_{k\geq 0, n\geq 0, k+n>0, \Delta\leq x} \frac{e^{-k-n}}{k!n!} e^{-2\lambda} \quad (1)$$

where $\Delta = \frac{(k-n)^2}{k+n}$, k, n are the natural numbers.

Below is the thus calculated distribution function Δ and the value χ_1^2 distribution function. It can be seen that the calculated distribution functions differ very little from the "chi-squared" -distribution even for a small λ.

Figure 1 Distribution Functions for Value Δ for λ=3, 10, 30 and for χ_1^2 distribution

Further increase of λ does not change the shape of the distribution, it only becomes smoother, close to continuous, the magnitude of the spikes decreases.

It results into a conclusion that to assess the probability of differences in incidence over time intervals, a sufficiently accurate estimate of the average incidence is not required, since the value λfor a not very small absolute incidence has little effect on the distribution under study.

Results

Data on the number of people who have been sick and died from COVID-19, according to the resource https://covid.observer/ were used for the analysis. Data on a relatively stable morbidity were selected for Russia for the period from July 1 to August 12, divided into data on Moscow, in which the incidence was quite high, however the considered period showed an

incidence of about 10 times lower than the maximum; and Russia exclusing Moscow, where the incidence either grew or remained at about the same level throughout the period of the entire year 2020. For comparison, data for Poland for the period from April 2 to May 27 as reference.

Figure 2 Number of cases and deaths from COVID-19 for Russia (separately: Moscow and the all other regions) for the period from 1 July to 12 August and Poland for the period from 2 April to 27 May

Analysis of the severity of differences in morbidity and mortality for Moscow are given in table №1. Table 1 shows that close values are too common for morbidity data for adjacent days. In particular, the values Δ that should have been present with the probability $F_\Delta < 0.1$ were observed in 11 cases out of 42, while the probability that the binomial distribution with $N = 42$ and $P = 0.1$ takes values of 11 or more is only 0.23%.

Table 1 The number of COVID-19 cases and deaths by day in Moscow from July 16 to August 12, the magnitude of the differences in incidence Δ for neighboring days and the probability F_Δ that such or a lower value may be accidental

There are also valid differences in the Δ distribution as well.

When analyzing data on Russia (exclusing Moscow) and Poland, we see the following

Table 2 Characteristics of the number of COVID-19 cases and deaths over the periods under review

The distributions of observed values, as has already been noted, differ from the expected uniform distribution, but the nature of differences is not the same:

Figure 3 Empirical Distributions of Value F_Δ for Moscow, Russia (excluding Moscow) and Poland

From fig. 3 it follows that for Moscow, both for the number of cases and for the number of deaths, small values of Δ are more often than expected, while for Russia excluding Moscow and Poland, large values of Δ are more often than expected.

It follows from table 2 that:

- For the considered period of relatively stable morbidity, the median of the number of cases of COVID-19 per day in Moscow was 671 cases, the median of incidence differences Δis 0.163, corresponding to F_Δ =0.313. For the number of deaths from COVID-19 per day we have a median of 14 cases, the median of differences for neighboring days Δ is 0.154, which corresponds to F_Δ=0.305. That means, for both the number of cases and the number of deaths, close values for neighboring days were more frequent than expected,

- If one conducts a similar analysis for Russia except Moscow, the median of the number of cases was 5240, the median of the differences for neighboring days Δ was 0.587, which corresponds to F_Δ=0.555. For the number of death, the median is 127, the median Δ is 2.777, corresponding to F_Δ=0.904.

If accepted assumptions correspond to the truth, the value F_Δ, like any distribution function, must be evenly distributed. When comparing the obtained distributions with uniform using the Kolmogorov-Smirnov criterion, it can be seen (table 2) that for morbidity in Moscow the difference is true with p=0.006, for Russia excluding Moscow differences are unreliable, p=0.126. At the same time, for mortality in both Moscow and Russia without Moscow, the differences in the actual distribution from the expected are true with p<0.001.

At the same time, for Poland, the differences in both the number of cases and the number of deaths in neighboring days are higher than expected with p<0.001.

Discussion

When analyzing the infectious morbidity, one of the characteristics is a focality, that is, the degree of grouping of individual cases, which can be a consequence of family hotbeds of disease, foci in organized children's groups, etc. So, if cases are detected not independently, but by N cases at once, it increases the incidence by N times, and the variance by N^2 times, that is, the ratio between the variance of the number cases and the number of cases can give an estimate about the size a foci.

The assumptions about the independence of individual cases are not entirely accurate, as both the causes of disease and their identification extend the effect not on one person, but on a group of individuals. This is especially pronounced for infectious diseases, as the emergence of a source of the pathogen increases the risk of disease for many contacts, and the detection of one case leads to more active identification among those in contact with him/her. Identifying one case increases the likelihood of detecting other cases, so the variance in the number of cases of disease should be greater than the mathematical expectation.

Also, the variance of morbidity might be increased by changes in the conditions of the epidemic process over time, in which the mathematical expectation of the number of cases and deaths varies over days.

Conclusion

The change in the conditions of the epidemic process, detection of cases, the criteria for making a pristine and posthumous diagnosis, and the fact that for infectious diseases individual cases are not are independent increase the incidence differences over adjacent time intervals. Therefore, in the analysis of the actual data, fluctuations of incidence are expected, greater than according to the distribution (1).

In the analysis of the actual incidence of COVID-19 in Poland, an increase in the difference in the number of cases and deaths in neighboring days was found with $p<0.001$. However, for data on the number of cases and deaths in Moscow, on the contrary, the difference in the number of cases and deaths in neighboring days is less than expected showing $p<0.001$, whereas for the data for Russia, with the exception of Moscow, no reliable differences from those obtained under the assumption of constancy and independence of cases of the disease were revealed.

It follows that at least for Moscow there is a deliberate smoothing of actual morbidity and mortality data, perhaps to reassure the population.

A few days after finishing work on the preliminary text of the article, one of the authors briefly mentioned on his blog that there are signs of manipulation of the data on morbidity and mortality from COVID-19 - the difference in the number of cases for adjacent days is too small. 3 days after that, starting from August 23, the differences in the number of cases of COVID-19 cases in Moscow over the next few days increased many times and began to correspond to the expected.

References

Abbott RJ, et al. 2017. Asymptomatic primary infection with Epstein-Barr virus: observations on young adult cases. *Journal of Virology* 91.21:1-49 DOI:10.1128/JVI.00382-17.

Black AJ, McKane AJ, Nunes A, Parisi A. 2009. Stochastic fluctuations in the susceptible-infective-recovered model with distributed infectious periods. *Physical Review E* 80.2 DOI: 10.1103/PhysRevE.80.021922.

Cooper GS, Bynum MLK, Somers EC. 2009. Recent insights in the epidemiology of autoimmune diseases: improved prevalence estimates and understanding of clustering of diseases. *Journal of Autoimmunity* 33.3-4: 197-207. DOI: 10.1016/j.jaut.2009.09.008.

Isanaka S, Boundy EO, Grais RF, Myatt M, Briend A. 2016. Improving Estimates of Numbers of Children With Severe Acute Malnutrition Using Cohort and Survey Data *American Journal of Epidemiology.* 184.2:861-869 DOI:10.1093/aje/kww129.

Krause AL, Kurowski L, Yawar K, Van Gorder RA. 2018. Stochastic epidemic metapopulation models on networks: SIS dynamics and control strategies. *Journal of Theoretical Biology* 449.14:35-52. DOI: 10.1016/j.jtbi.2018.04.023.

Leung KY, Trapman P, Britton T. 2018. Asymptomatic Primary Infection with Epstein-Barr Virus: Observations on Young Adult Cases. *Mathematical Biosciences* 301:190-198 DOI: 10.1016/j.mbs.2018.04.002.

Simões M, Telo da Gama M M, Nunes A. 2008. Stochastic fluctuations in epidemics on networks. *Journal of The Royal Society Interface* 5.22: 555-566 DOI: 10.1098/rsif.2007.1206.

Stuart A., Ord K. 2009. *Kendall's Advanced Theory of Statistics, Distribution Theory 6th Edition*. NY, Wiley, 700.

Yu WY, Zhang QQ, Xiao Y, Tan WG, Li XD, Lu PX. 2019. Correlation between 18F-FDG PET CT SUV and symptomatic or asymptomatic pulmonary tuberculosis. *Journal of X-ray Science and Technology* 27.5:899-906 DOI: 10.3233/xst-190543.

Герасимов А.Н., Михеева И.В. 2018. Эпидемиологическая ситуация в России – кажущееся благополучие и скрытые угрозы. *Тихоокеанский медицинский журнал* 73.3:74-78 DOI: 10.17238/PmJ1609-1175.2018.3.75-78.

Герасимов А.Н. 2014. Определение статистической достоверности различий в заболеваемости. *Эпидемиология и Вакцинопрофилактика* 78.5:54-60:. https://www.elibrary.ru/download/elibrary_22416303_93752572.pdf

Table 1(on next page)

The number of COVID-19 cases and deaths by day in Moscow from July 15 to August 12, the magnitude of the differences in incidence Δ for neighboring days and the probability F_{Δ} that such or a lower value may be accidental

Table 1

The number of COVID-19 cases and deaths by day in Moscow from July 15 to August 12, the magnitude of the differences in incidence Δ for neighboring days and the probability F_Δ that such or a lower value may be accidental

Date	Number of new cases per day		Δ		F_Δ(), the exact solution according to formula (1)		F_Δ (), approximate solution according to χ^2_1 destribution	
	Ill	Dead	Ill	Dead	Ill	Dead	Ill	Dead
16.07.2020	531	24	8,118	0,472	0,9956	0,5106	0,9956	0,9956
17.07.2020	575	13	1,750	3,270	0,8143	0,9303	0,8142	0,8142
18.07.2020	578	14	0,008	0,037	0,0705	0,1569	0,0704	0,0704
19.07.2020	591	14	0,145	0,000	0,2963	0,0757	0,2962	0,2962
20.07.2020	578	15	0,145	0,034	0,2963	0,1513	0,2962	0,2962
21.07.2020	602	17	0,488	0,125	0,5154	0,2803	0,5152	0,5152
22.07.2020	638	19	1,045	0,111	0,6935	0,2646	0,6934	0,6934
23.07.2020	608	14	0,722	0,758	0,6047	0,6200	0,6046	0,6046
24.07.2020	645	11	1,093	0,360	0,7042	0,4573	0,7041	0,7041
25.07.2020	648	14	0,007	0,360	0,0666	0,4573	0,0665	0,0665
26.07.2020	683	9	0,920	1,087	0,6627	0,7070	0,6626	0,6626
27.07.2020	694	13	0,088	0,727	0,2332	0,6127	0,2331	0,2331
28.07.2020	674	10	0,292	0,391	0,4114	0,4744	0,4113	0,4113
29.07.2020	671	13	0,007	0,391	0,0653	0,4744	0,0652	0,0652
30.07.2020	678	12	0,036	0,040	0,1512	0,1631	0,1512	0,1512
31.07.2020	695	14	0,210	0,154	0,3537	0,3102	0,3536	0,3536
01.08.2020	690	13	0,018	0,037	0,1070	0,1569	0,1069	0,1069
02.08.2020	664	12	0,499	0,040	0,5203	0,1631	0,5202	0,5202
03.08.2020	693	13	0,620	0,040	0,5690	0,1631	0,5689	0,5689
04.08.2020	691	12	0,003	0,040	0,0430	0,1631	0,0429	0,0429
05.08.2020	687	11	0,012	0,043	0,0859	0,1702	0,0858	0,0858
06.08.2020	684	13	0,007	0,167	0,0647	0,3224	0,0646	0,0646
07.08.2020	686	12	0,003	0,040	0,0432	0,1631	0,0431	0,0431
08.08.2020	691	14	0,018	0,154	0,1073	0,1631	0,1072	0,1072
09.08.2020	689	12	0,003	0,154	0,0430	0,1631	0,0429	0,0429
10.08.2020	694	13	0,018	0,040	0,1070	0,1631	0,1070	0,1070
11.08.2020	694	14	0,000	0,037	0,0107	0,1569	0,0000	0,0000
12.08.2020	689	12	0,018	0,154	0,1070	0,1631	0,1070	0,1070

5

Table 2(on next page)

Characteristics of the number of COVID-19 cases and deaths over the periods under review

Table 2

Characteristics of the number of COVID-19 cases and deaths over the periods under review

	Moscow		Russia excluding Moscow		Poland	
	Ill	Dead	Ill	Dead	Ill	Dead
Increase rate per day,%	0.26%	-2.91%	-0.89%	-0.97%	-0.06%	-0.88%
p (comparison of F_Δ distribution with uniform, Kolmogorov-Smirnov criterion)	0.006	<0.001	0.127	<0.001	<0.001	<0.001
Median						
Number of cases per day	671	14	5240	127	337	13
Δ	0.163	0.154	0.587	2.777	5.075	1.000
F_Δ	0.313	0.250	0.555	0.904	0.976	0.701

Figure 1

Distribution Functions for Value Δ for λ=3, 10, 30 and for χ^2 distribution

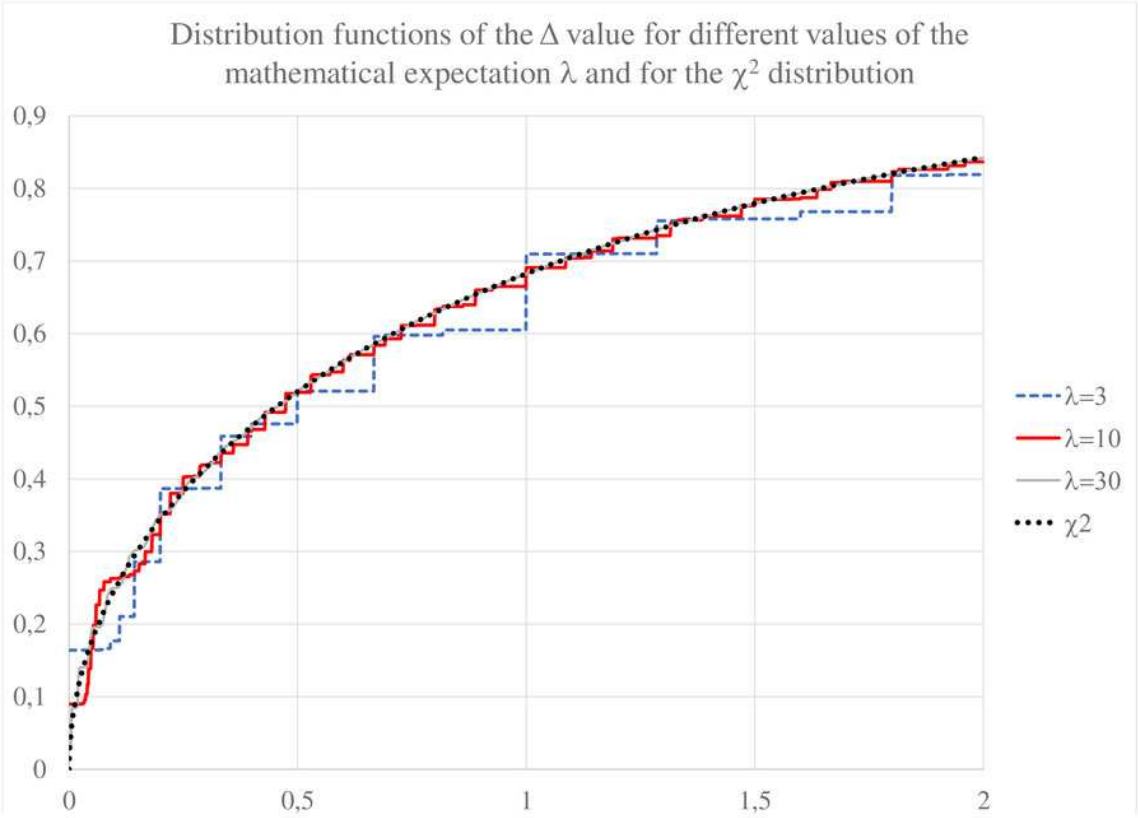

Figure 2

Number of cases and deaths from COVID-19 for Russia (separately: Moscow and the all other regions) for the period from 1 July to 12 August and Poland for the period from 2 April to 27 May

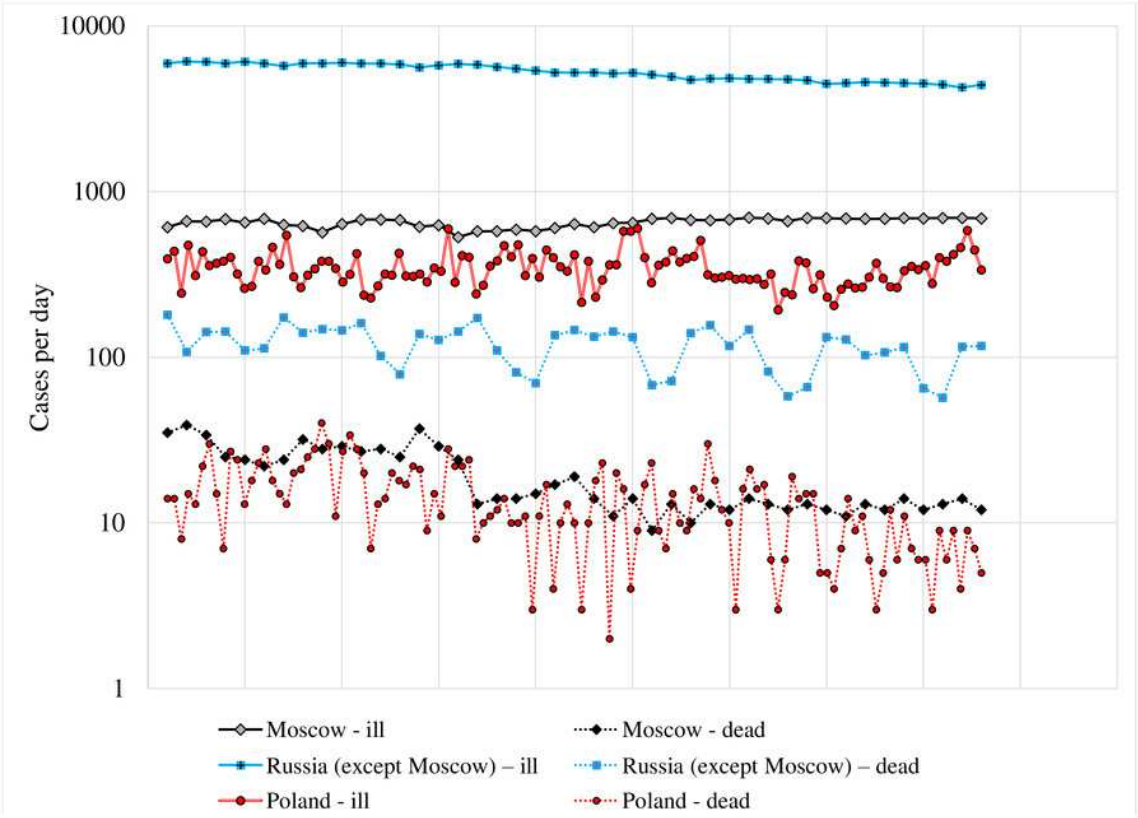

Figure 3

Empirical Distributions of Value F_{Δ} for Moscow, Russia (excluding Moscow) and Poland