Infinite variance of U.S. COVID-19 cases & deaths, & Taylor's law of heavy-tailed data

Joel E. Cohen, cohen@rockefeller.edu

Rockefeller University & Columbia University

ICTP Workshop on Quantitative Human Ecology, Trieste 2022-07-28

Ashokan Reservoir, Ulster County, NY, JEC 20151015

Plan

Fluctuation scaling, variance function

$variance = f(mean)$

Taylor's law: $variance = a(mean)^b$

$\log(variance) = \log a + b \log(mean)$

Heavy tails & regular variation

$\Pr(X > x) = L(x)x^{-\alpha}, 0 < \alpha < 2$

COVID-19 in US

Taylor's law, infinite variance

Plan

→Fluctuation scaling, variance function

$variance = f(mean)$

Taylor's law: $variance = a(mean)^b$

$\log(variance) = \log a + b\log(mean)$

Heavy tails & regular variation

$\Pr(X > x) = L(x)x^{-\alpha}, 0 < \alpha < 2$

COVID-19 in US

Taylor's law, infinite variance

Variance function

Population: Given a non-empty family of random variables $\{X(s)\}_{s \in S}$, if each $X(s)$ has finite mean $E(X(s))$ & finite variance $Var(X(s))$, the population variance function f says: $Var\big(X(s)\big) = f(E[X(s)]), \forall s \in S$.

Sample: Sample of size $n > 1$ is a set $\{X_1(s), \ldots, X_n(s)\}$ of n iid copies of $X(s)$, with sample mean $\bar{X}_n(s) \coloneqq (X_1(s) + \cdots + X_n(s))/n$, sample variance $s_n^2(s)$. The sample variance function f_n says: $s_n^2(s) \approx f_n(\bar{X}_n(s))$.

Plan

Fluctuation scaling, variance function

$variance = f(mean)$

→Taylor's law: $variance = a(mean)^b$

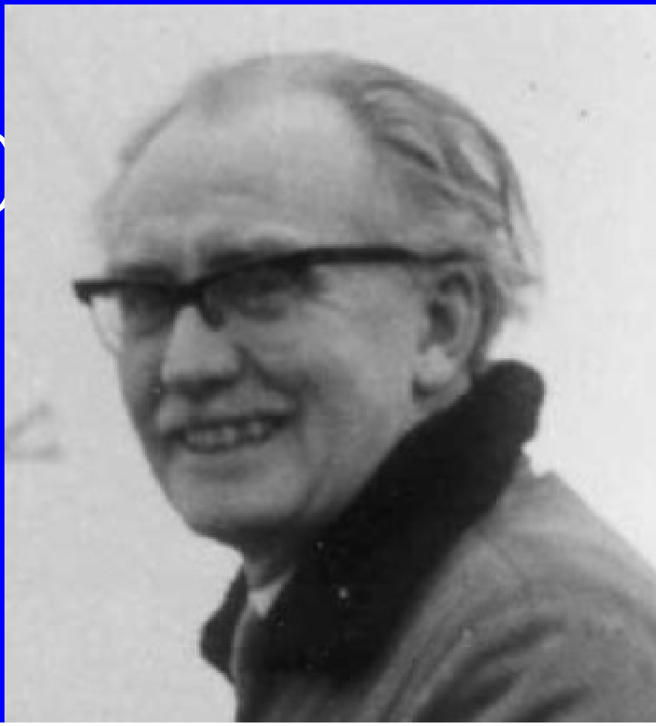
$\log(variance) = \log a + b \log(mean)$

Heavy tails & regular variation

$\Pr(X > x) = L(x)x^{-\alpha}, \alpha \in (0,2)$

COVID-19 in US

Taylor's law, infinite variance



Lionel Roy Taylor
(1924–2007)

Taylor's law(s)

Population TL holds if each nonnegative random variable $X(s), \forall s \in S$, has finite, positive mean & variance & $\exists\ a > 0, b$ such that

$$Var\big(X(s)\big) = a\{E\big(X(s)\big)\}^b.$$

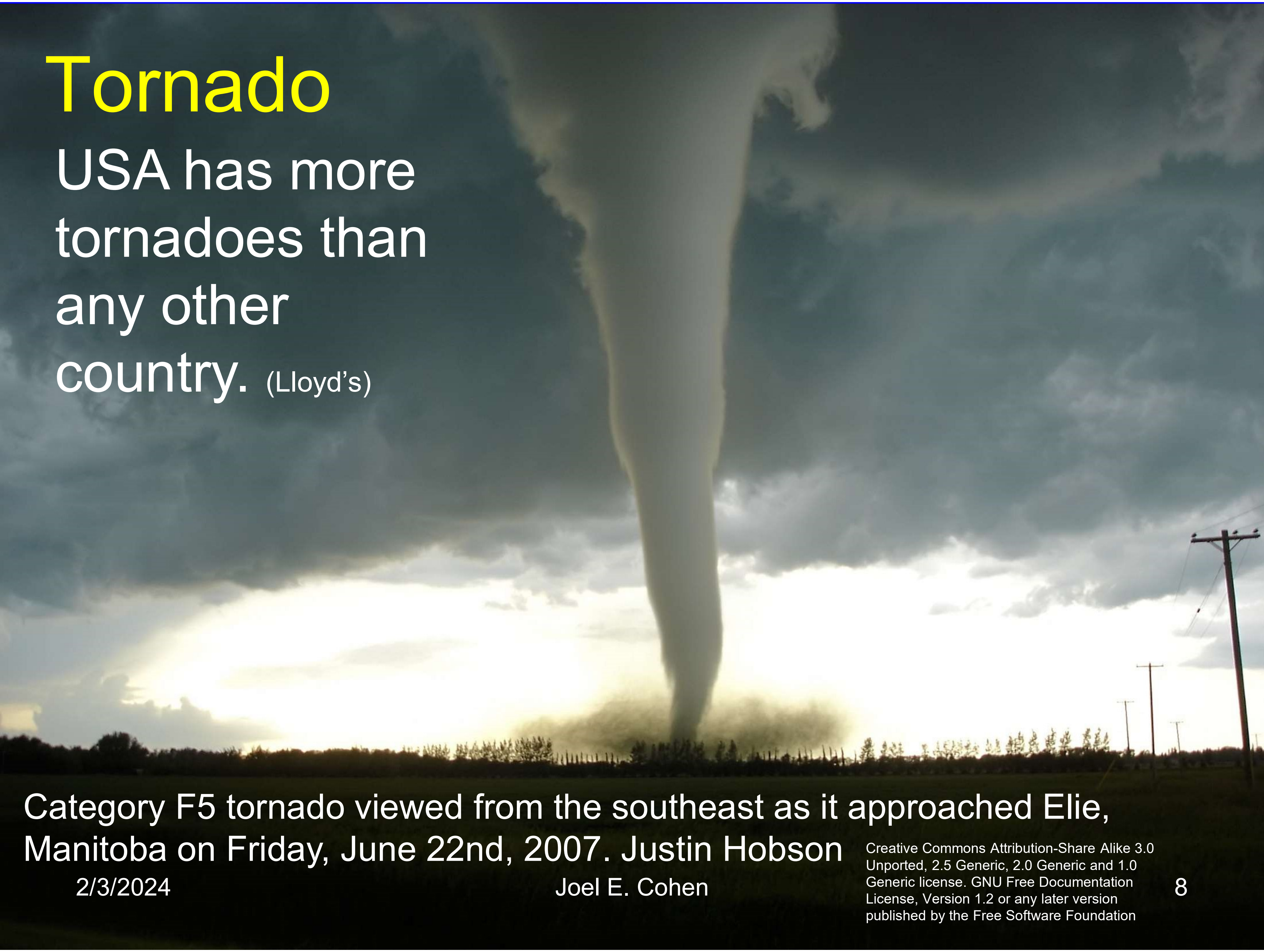
Sample TL holds if samples with mean $\bar{X}_n(s)$, variance $s_n^2(s)$ obey, $\forall s \in S$, for some $a > 0, b$, $\log s_n^2(s) \approx \log a + b \log \bar{X}_n(s)$ or

$$\frac{s_n^2(s)}{\{\bar{X}_n(s)\}^b} \approx a > 0.$$

For sample TL, there is no requirement that mean or variance exist or are finite.

TL data structure: multiple samples, each with multiple observations

Sample number →	*s*=1	*s*=2	*s*=3	*s*=…
Population size or density in units (quadrats, plots, transects, counties, states, years, days)	x_{11}	x_{12}	x_{13}	$x_{\ldots}$
	x_{21}	x_{22}	x_{23}	…
	x_{31}	x_{32}	x_{33}	…
		x_{42}	x_{43}	…
		x_{52}		…
Mean (weighted)	m_1	m_2	m_3	m…
Variance (weighted)	v_1	v_2	v_3	v…



Tornado

USA has more tornadoes than any other country. (Lloyd's)

Category F5 tornado viewed from the southeast as it approached Elie, Manitoba on Friday, June 22nd, 2007. Justin Hobson

TL for tornadoes: size of outbreaks by calendar year

Year→	1954	1955	…	2014
Number of F1+ tornadoes per outbreak (≥6 consecutive tornadoes with ≤6 hour gap)	x_{11}	x_{12}	x_{13}	$x_{\dots}$
	x_{21}	x_{22}	x_{23}	…
	x_{31}	x_{32}	x_{33}	…
		x_{42}	x_{43}	…
		x_{52}		…
Mean (weighted)	m_1	m_2	m_3	m…
Variance (weighted)	v_1	v_2	v_3	v…

Outbreaks (≥6 tornadoes) cause most damage.

Outbreak is defined as ≥6 tornadoes starting ≤6 hours apart.

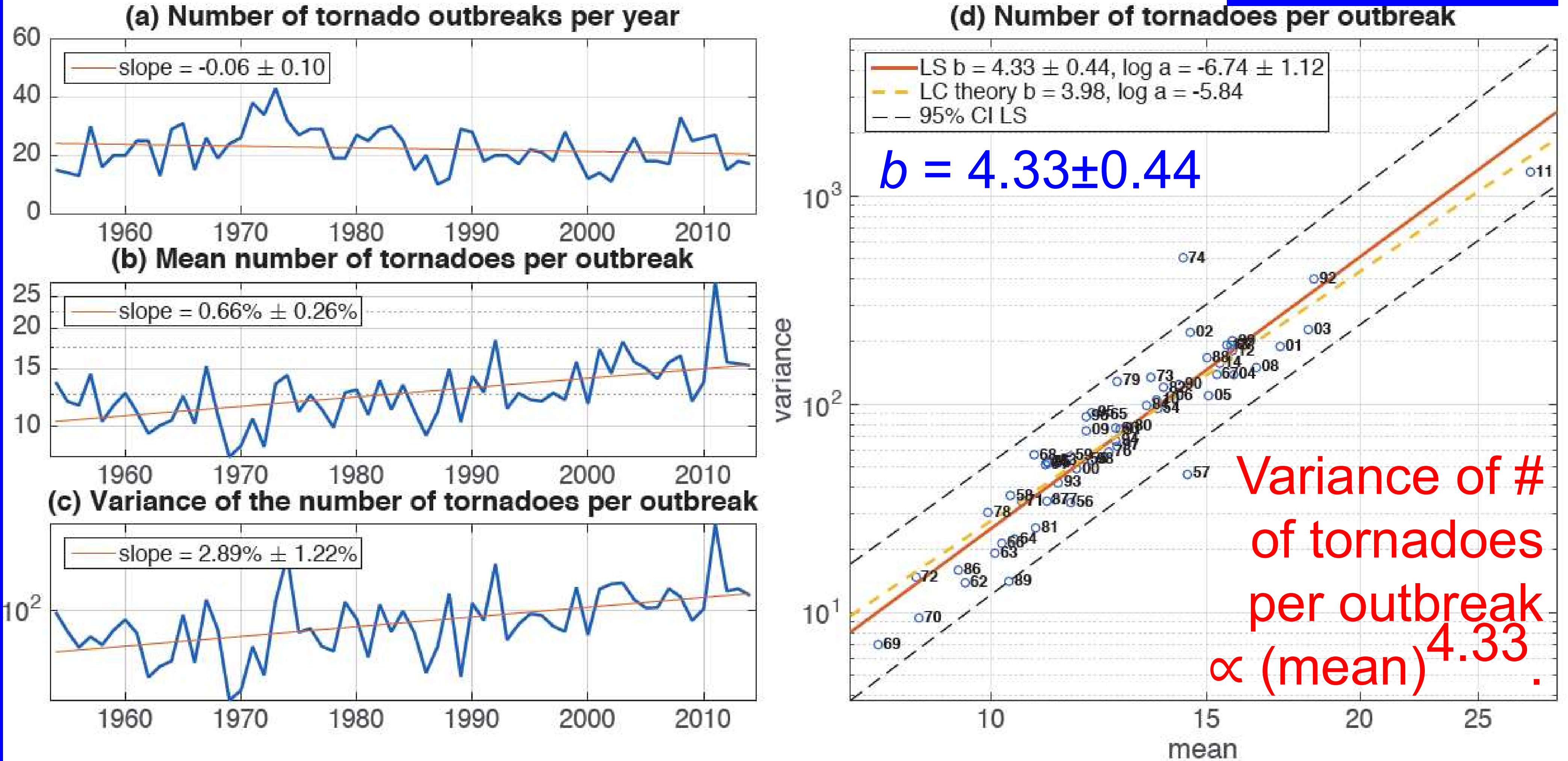
1972–2010: 79% of tornado fatalities & most economic losses occurred in outbreaks.

No trends in numbers of reliably reported tornadoes or outbreaks in last half century.

Mean & variance of the number of tornadoes per outbreak, & insured losses, increased significantly in last half century.

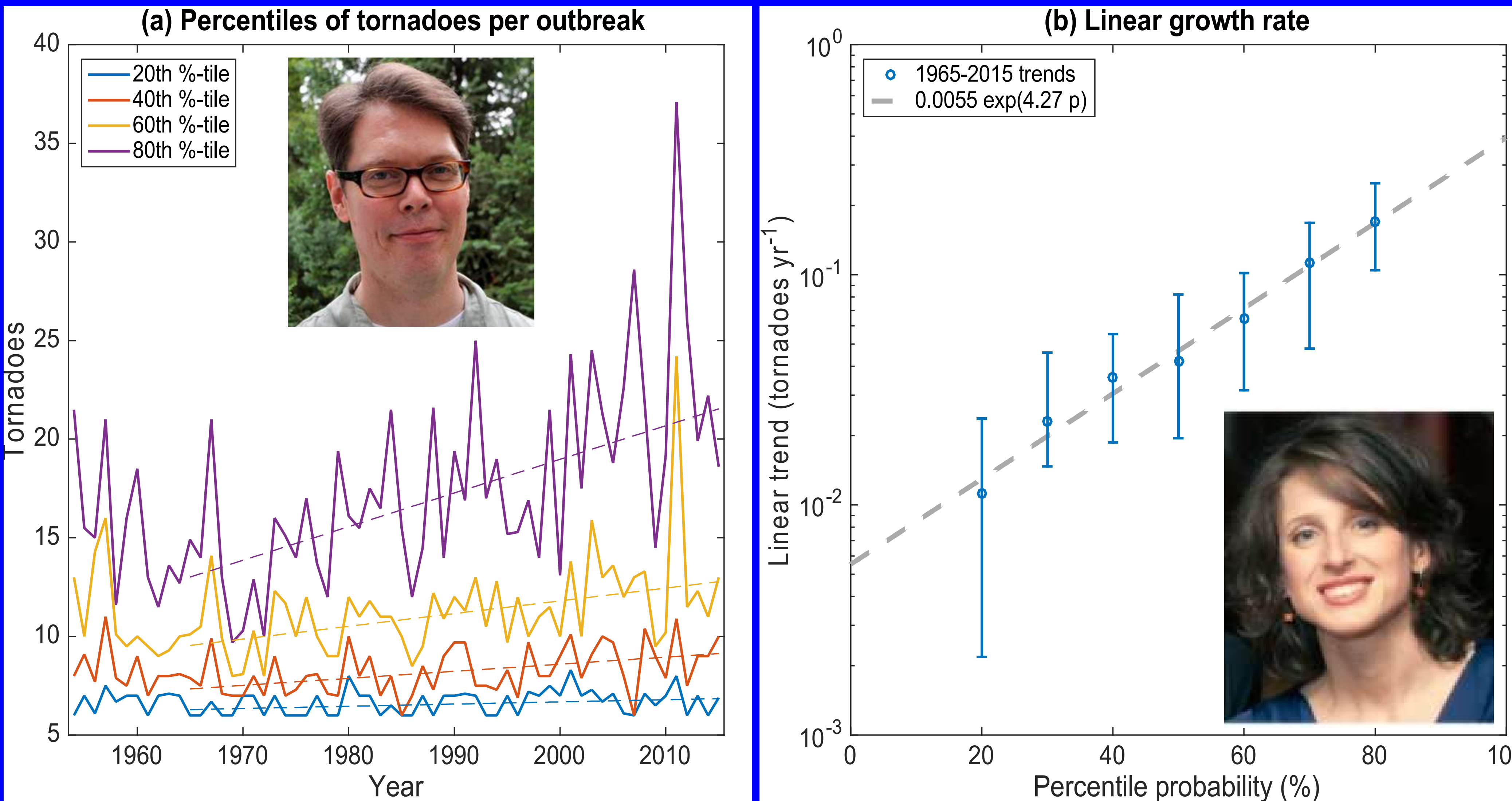
F1+ tornadoes per outbreak in USA: variance~(mean)$^{4.3}$

Tippett & Cohen, *Nature Communications* 2016



Higher percentiles increased faster.

"quantile regression"



Tippett, Lepore, & Cohen, *Science* 2016

Plan

Fluctuation scaling, variance function

$variance = f(mean)$

Taylor's law: $variance = a(mean)^b$

$\log(variance) = \log a + b \log(mean)$

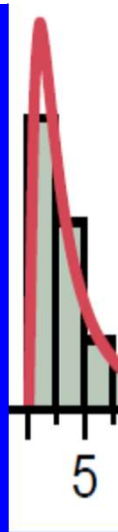
→Heavy tails & regular variation

$\Pr(X > x) = L(x)x^{-\alpha}, 0 < \alpha < 2$

COVID-19 in US

Taylor's law, infinite variance

Lognormal distribution



probability density function of lognormal(1,1)

A positive-valued random variable $Y(\mu, \sigma^2)$ with real parameters $\mu, \sigma^2 \geq 0$ is ***lognormal*** if $\log Y(\mu, \sigma^2)$ is normal(mean μ, variance σ^2).

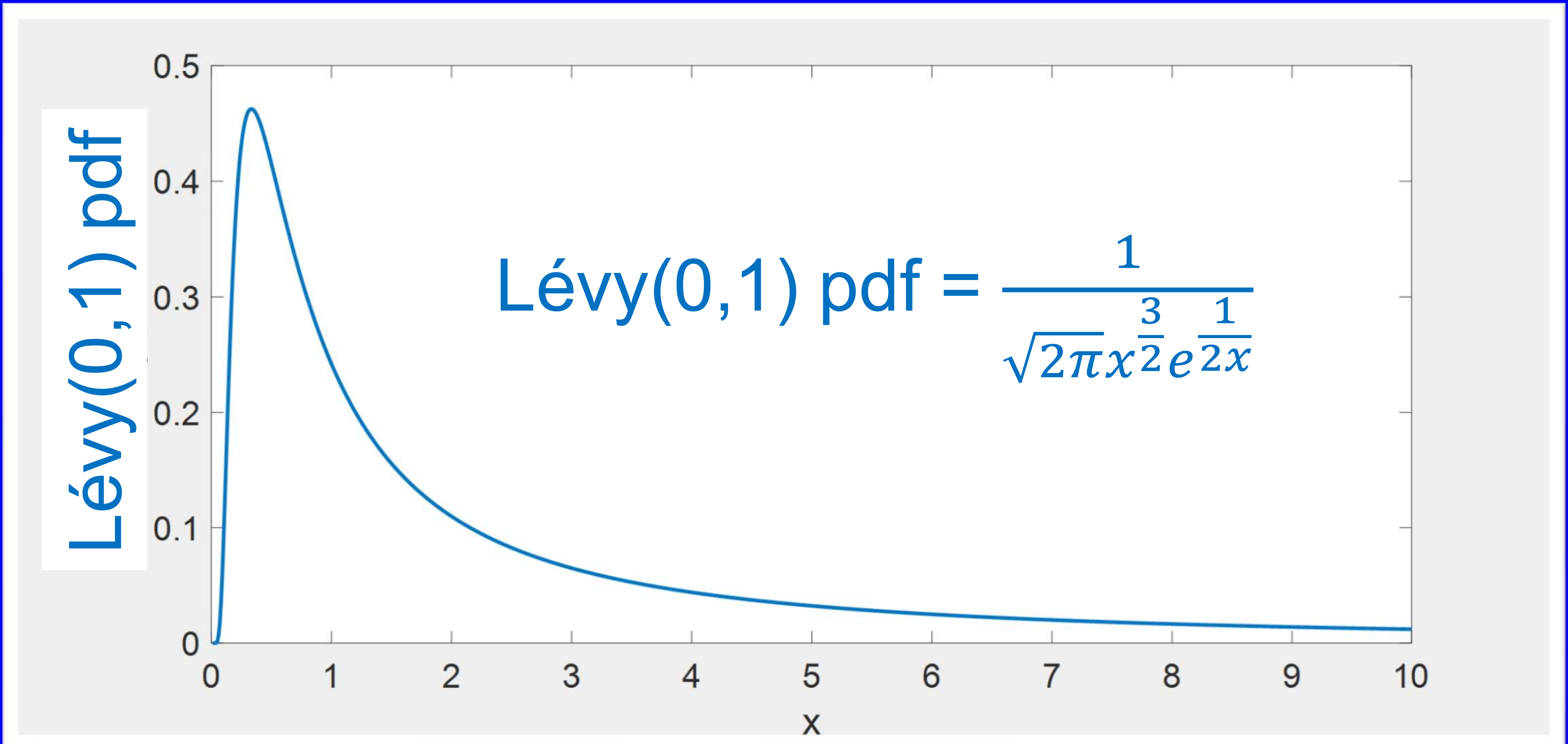
$$E\left(Y(\mu, \sigma^2)\right) = \exp\left(\mu + \frac{\sigma^2}{2}\right),$$

$$\mathrm{Var}\left(Y(\mu, \sigma^2)\right) = [\exp(\sigma^2) - 1] \exp(2\mu + \sigma^2).$$

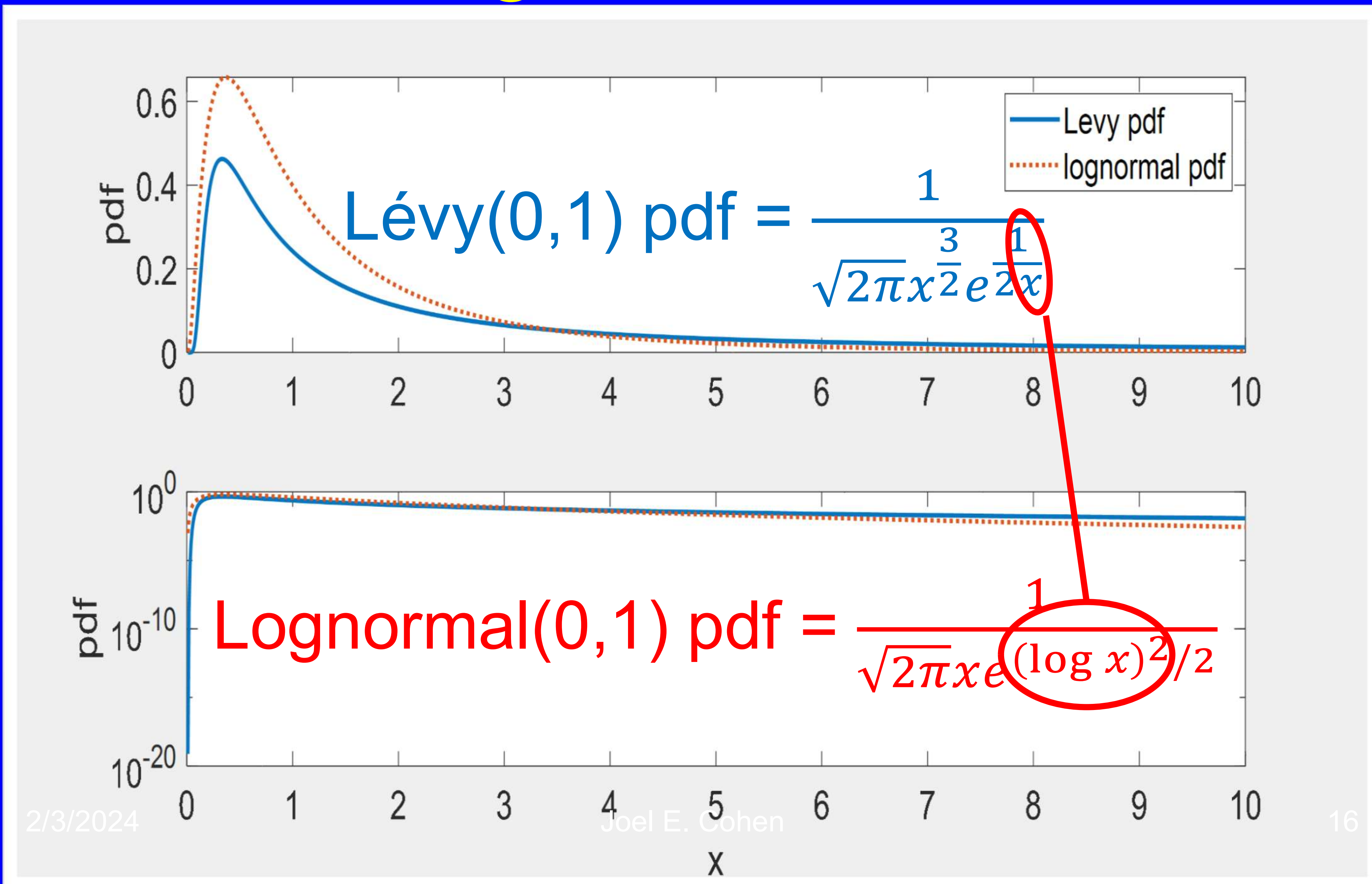
If σ^2 is constant & only μ changes, then $\mathrm{Var}(Y(\mu)) = c\left\{E\left(Y(\mu)\right)\right\}^2$: TL with exponent 2.

Lévy distribution $S_{1/2}$

If X is normal $\mathcal{N}(0,1)$, then $1/X^2$ has Lévy distribution (1924) $S_{1/2}$ [Helmert 1875, Lüroth 1876] with infinite mean & variance.



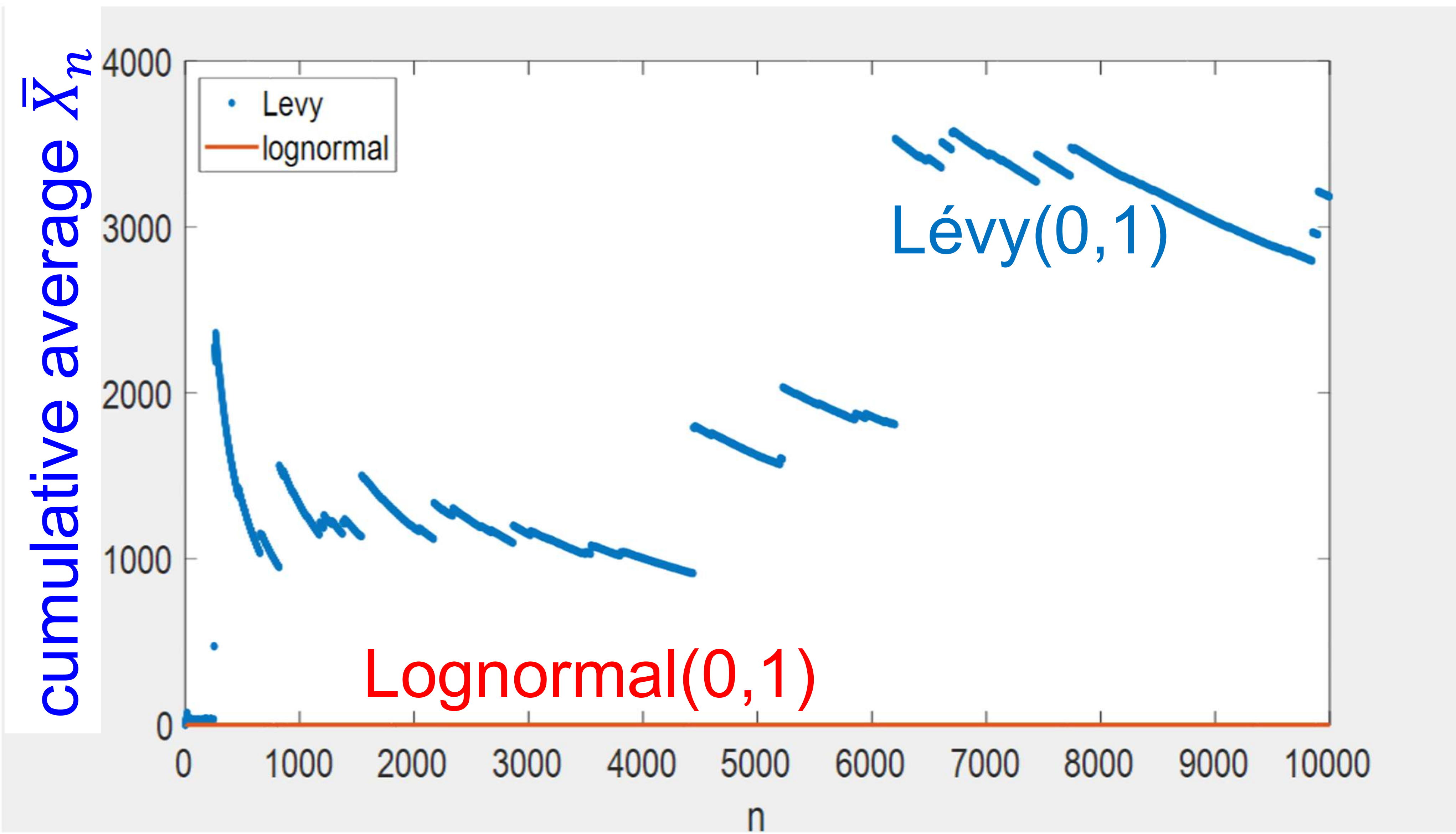
Lévy distribution has heavier right tail than lognormal distribution.



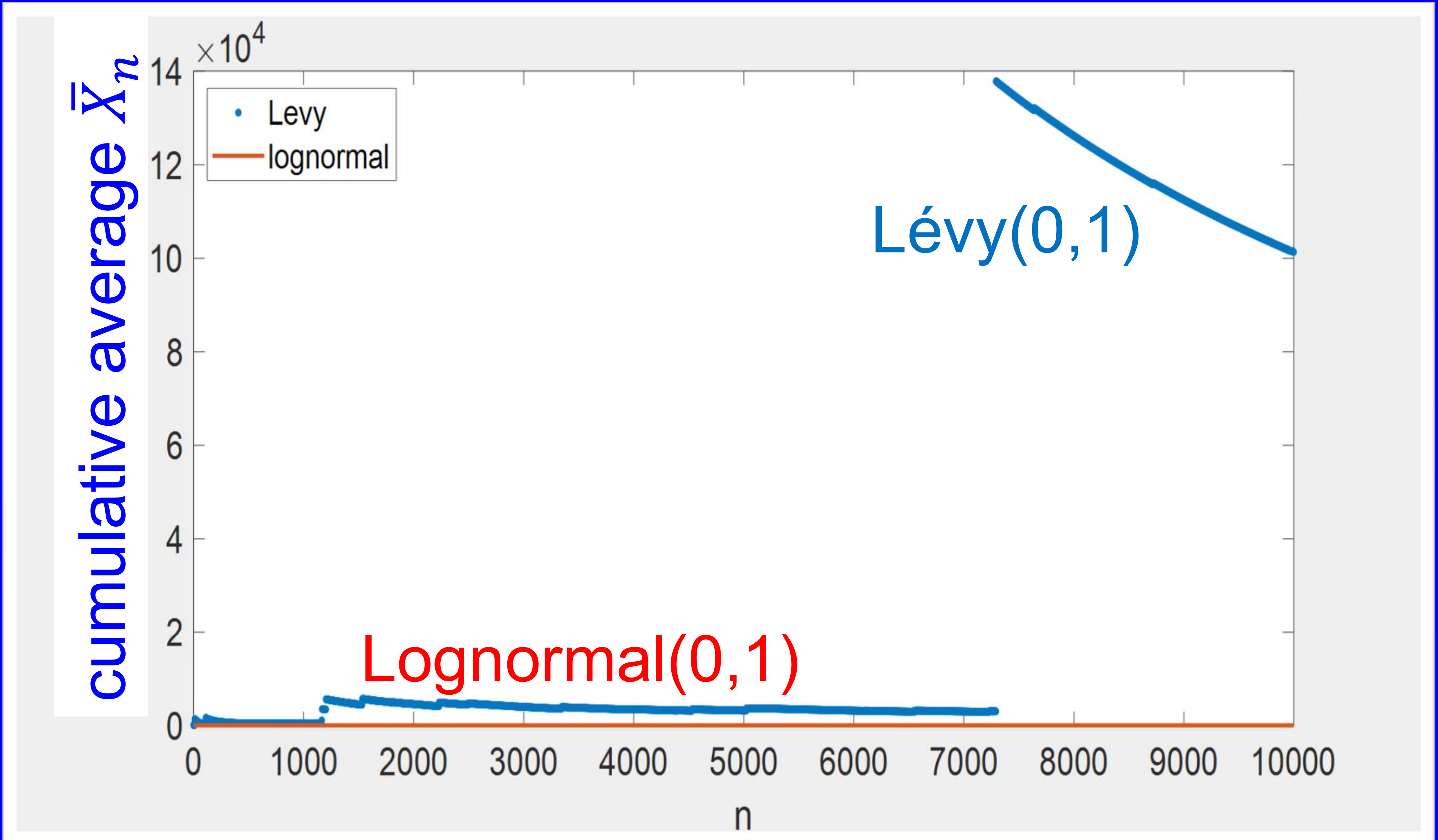
$$\text{Lévy(0,1) pdf} = \frac{1}{\sqrt{2\pi}x^{\frac{3}{2}}e^{\frac{1}{2x}}}$$

$$\text{Lognormal(0,1) pdf} = \frac{1}{\sqrt{2\pi}xe^{(\log x)^2/2}}$$

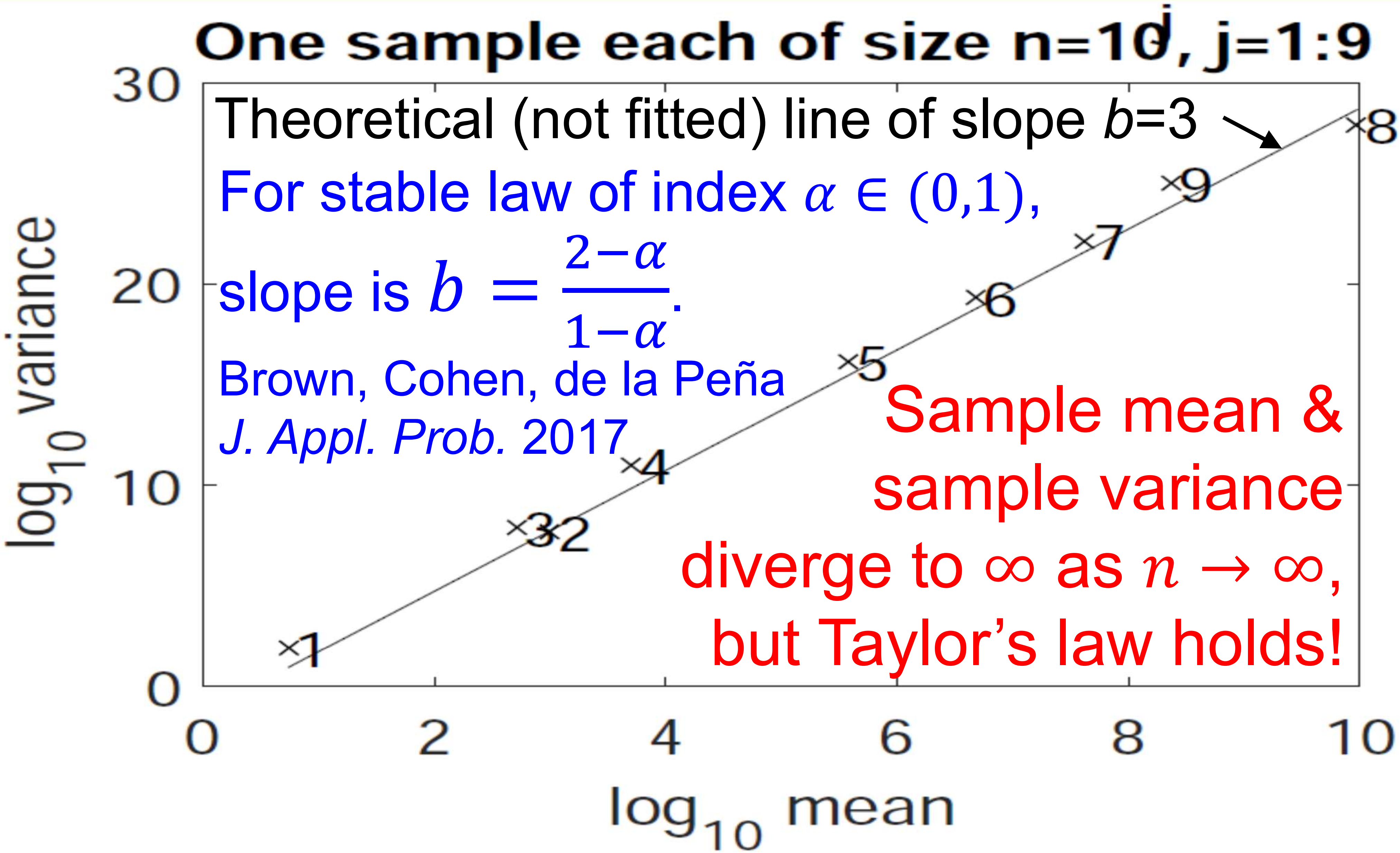
Lévy cumulative averages grow like $n \times$ Lévy. Lognormal averages converge.



Lévy cumulative averages grow like $n \times$ Lévy. Lognormal averages converge.



Lévy law (stable $\alpha = 1/2$) obeys TL with increasing sample sizes.



Heterogeneous dependent data

Cohen, Davis, Samorodnitsky, *Proc. Roy. Soc. A* 2020

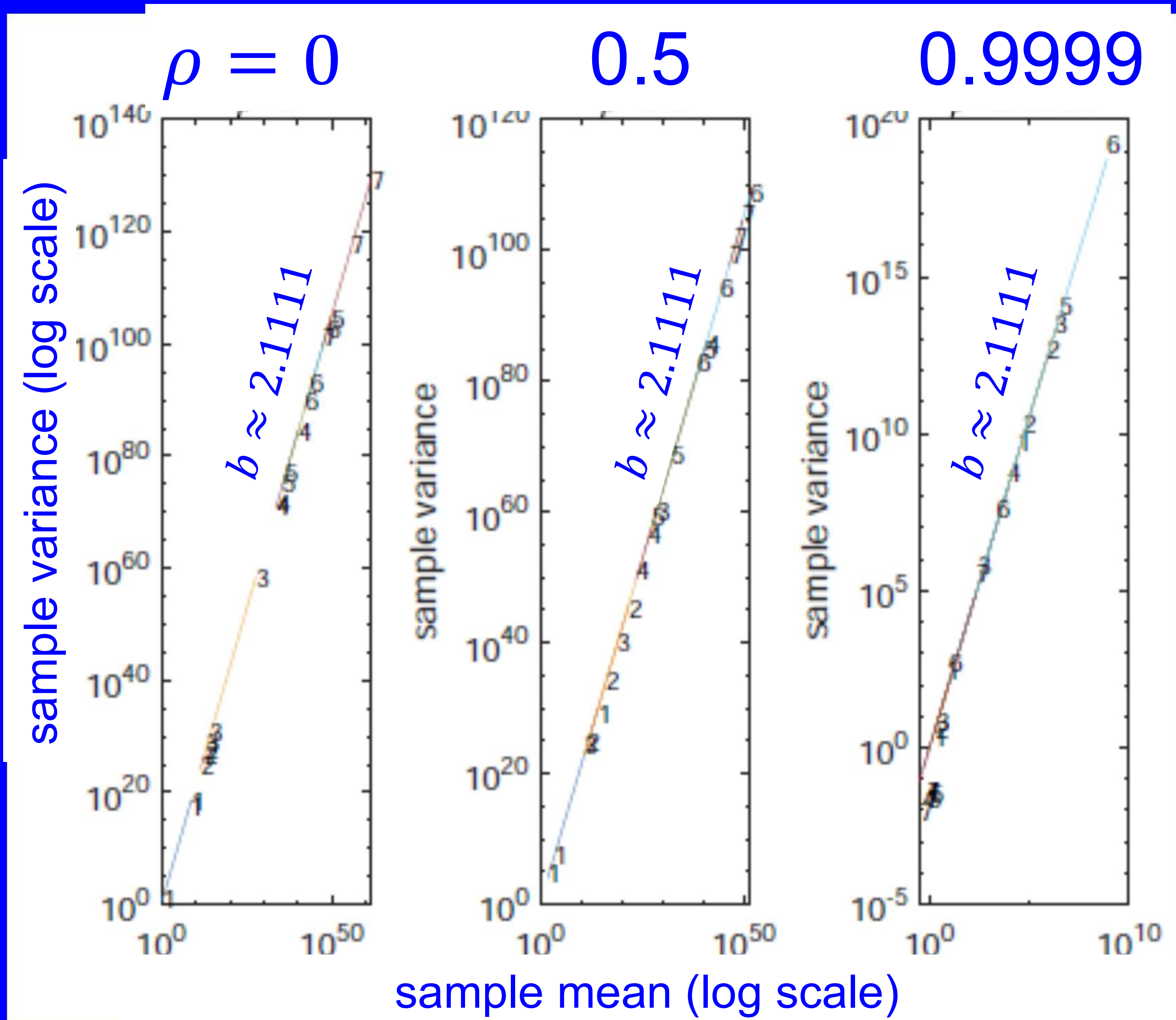
$$X_t := \frac{1}{|Z|^{\frac{1}{\alpha_t}}},$$

$$\Pr\{\alpha_t = 0.1\} = 0.1,$$

$$\Pr\{\alpha_t = 0.9\} = 0.9,$$

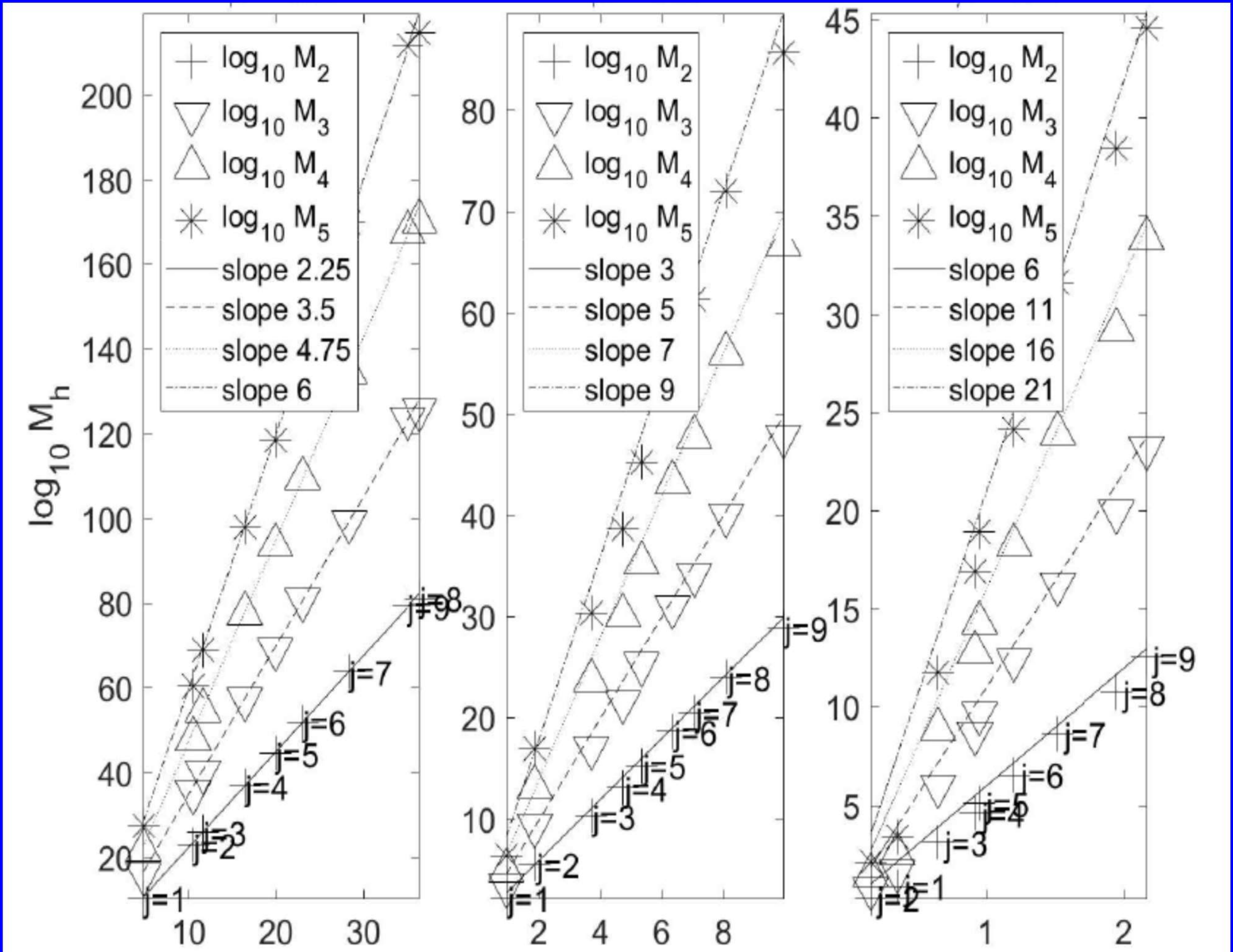
$$corr(Z_s, Z_t) = \rho, \text{ for } s \neq t.$$

$$b = \frac{2 - 0.1}{1 - 0.1} \approx 2.1111.$$



α=0.2 α=0.5 α=0.8

Log central moments



$\log(\bar{X}_n)$, $n = 10^j$

Many roads lead to TL.

Many models yield TL exactly or asymptotically.

Power-law form & parameter values of TL do not determine underlying mechanisms.

Interpreting the parameters of TL in terms of a specific mechanism requires testing the assumptions against detailed data.

Plan

Fluctuation scaling, variance function

$variance = f(mean)$

Taylor's law: $variance = a(mean)^b$

$\log(variance) = \log a + b \log(mean)$

Heavy tails & regular variation

$\Pr(X > x) = L(x)x^{-\alpha}, 0 < \alpha < 2$

→COVID-19 in US

Taylor's law, infinite variance

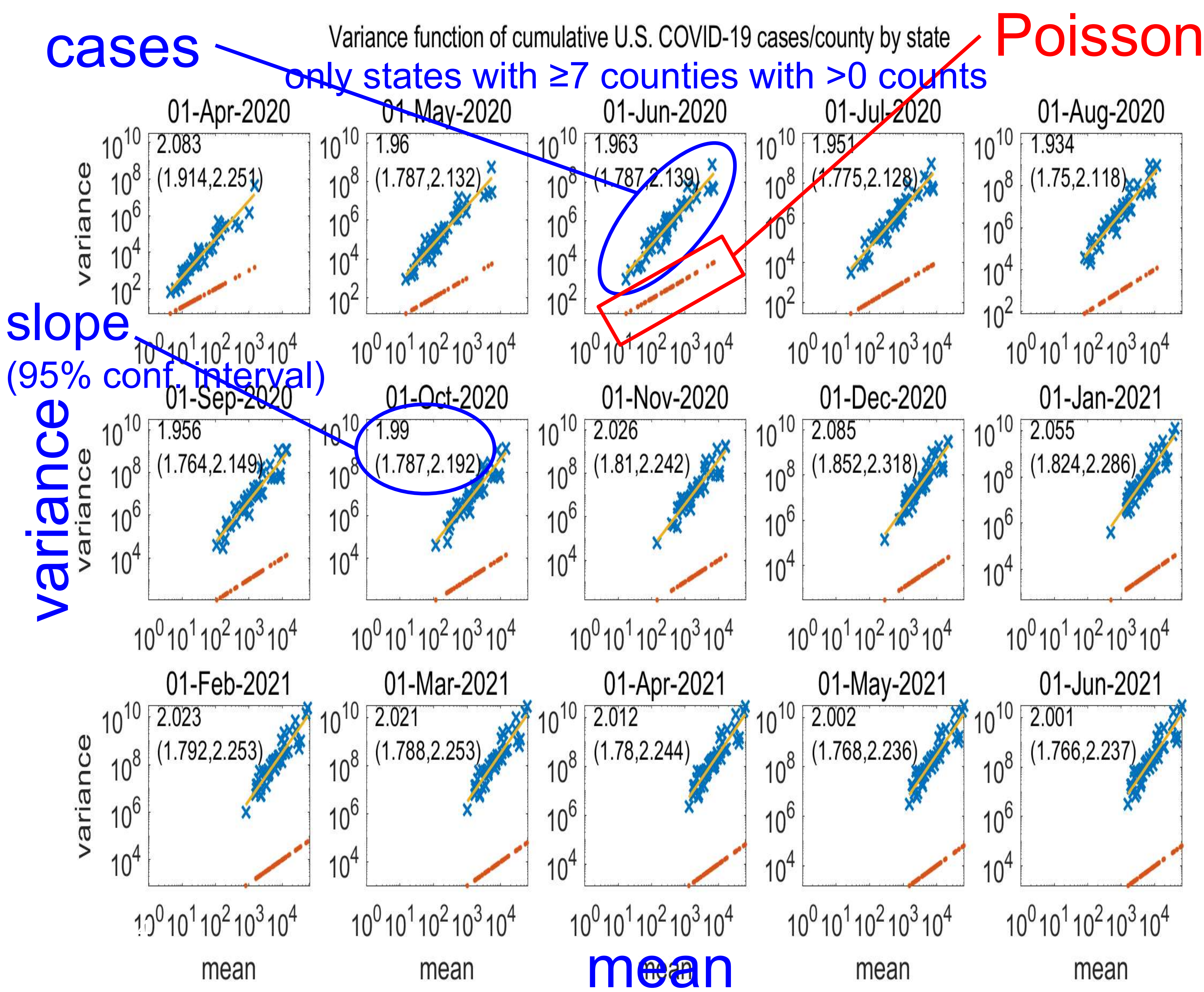
COVID-19 cases & deaths

New York Times historical data base has final counts of COVID-19 cumulative cases & cumulative deaths at end of each day, 2020-01-21 to 2021-06-19 by "state" & "county" for days & counties with >0 cases or >0 deaths.

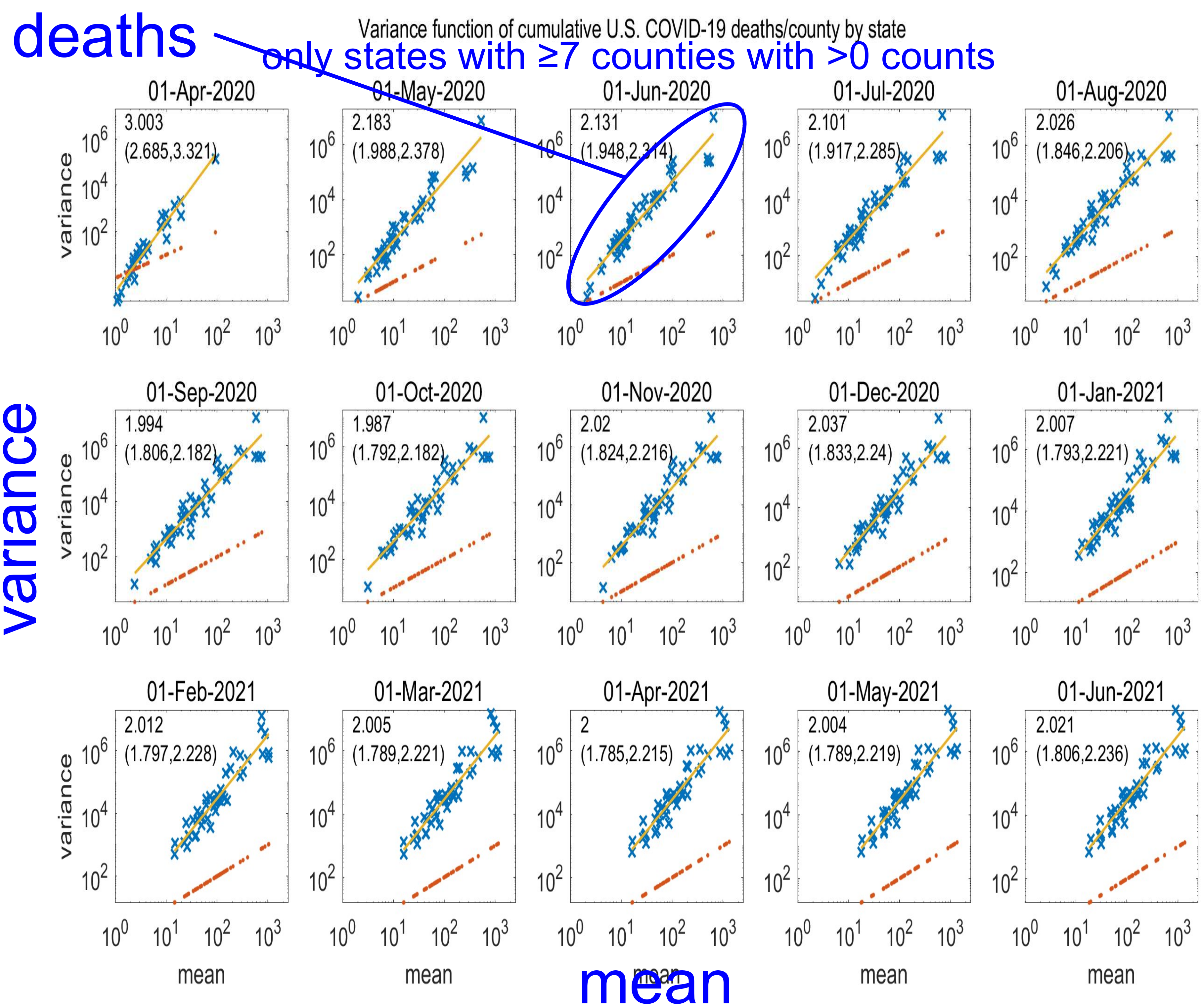
1,436,628 counts by day & county in data downloaded 2021-06-20

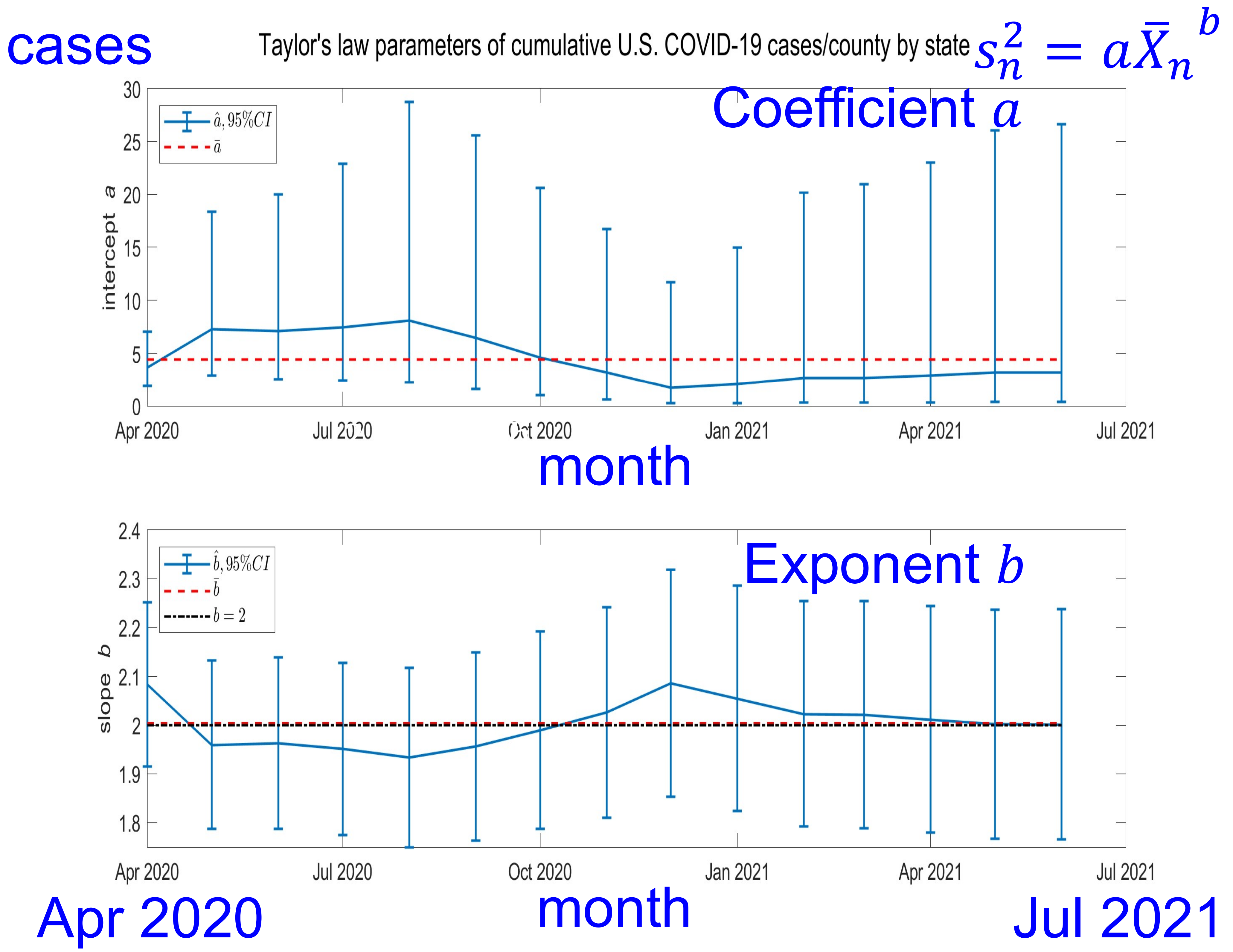
On each date, cumulative cases & deaths within each state by county

State number →	s=1	s=2	s=3	s=…
County 1	x_{11}	x_{12}	x_{13}	$x_{\dots}$
County 2	x_{21}	x_{22}	x_{23}	…
County 3	x_{31}	x_{32}	x_{33}	…
County 4		x_{42}	x_{43}	…
⋮		x_{52}		…
Mean = average	m_1	m_2	m_3	m…
Variance	v_1	v_2	v_3	v…

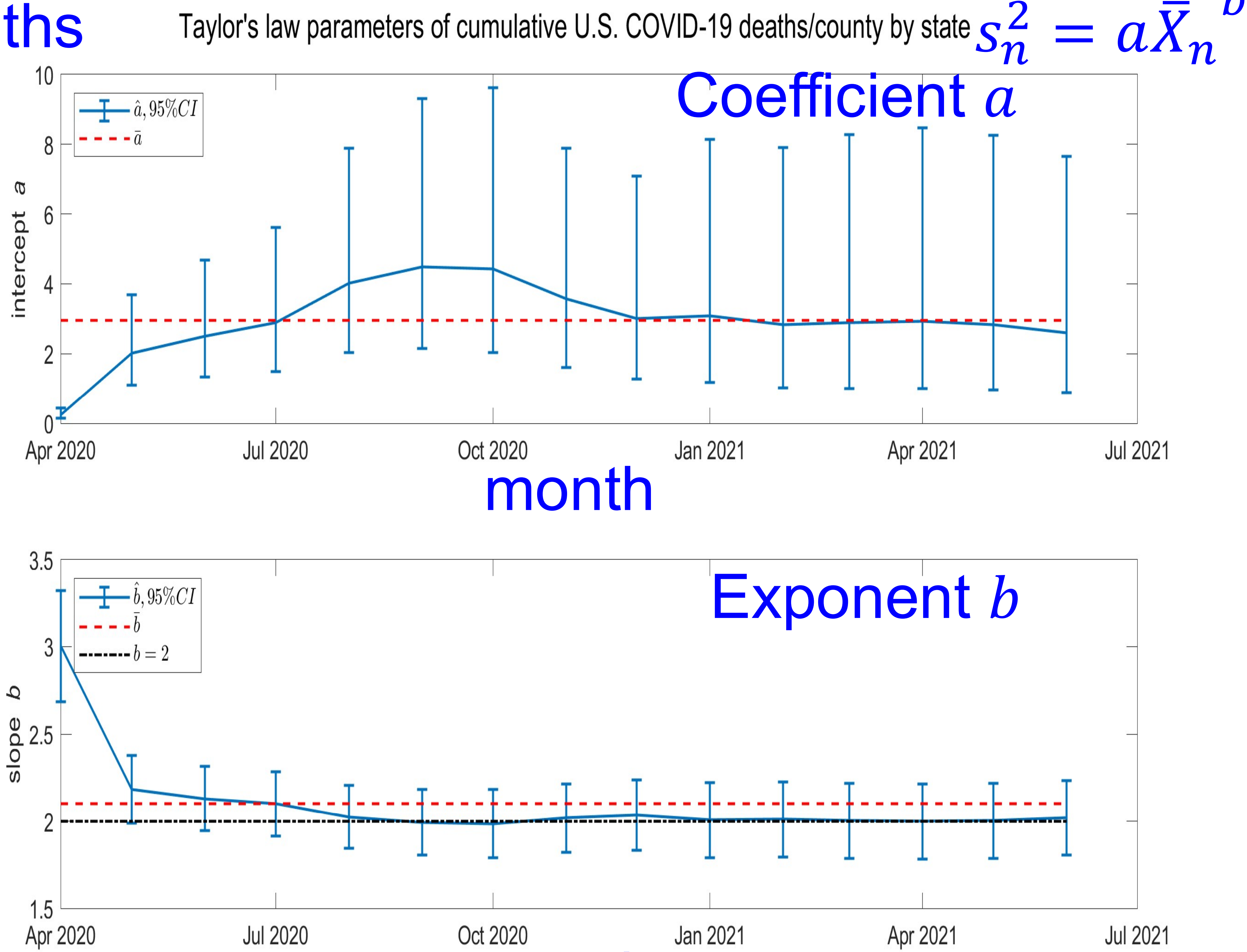


cases
Variance function of cumulative U.S. COVID-19 cases/county by state
only states with ≥7 counties with >0 counts
Poisson
slope
(95% conf. interval)
variance
mean
01-Apr-2020
2.083
(1.914,2.251)
01-May-2020
1.96
(1.787,2.132)
01-Jun-2020
1.963
(1.787,2.139)
01-Jul-2020
1.951
(1.775,2.128)
01-Aug-2020
1.934
(1.75,2.118)
01-Sep-2020
1.956
(1.764,2.149)
01-Oct-2020
1.99
(1.787,2.192)
01-Nov-2020
2.026
(1.81,2.242)
01-Dec-2020
2.085
(1.852,2.318)
01-Jan-2021
2.055
(1.824,2.286)
01-Feb-2021
2.023
(1.792,2.253)
01-Mar-2021
2.021
(1.788,2.253)
01-Apr-2021
2.012
(1.78,2.244)
01-May-2021
2.002
(1.768,2.236)
01-Jun-2021
2.001
(1.766,2.237)


deaths
Variance function of cumulative U.S. COVID-19 deaths/county by state
only states with ≥7 counties with >0 counts
variance
mean
01-Apr-2020
3.003
(2.685,3.321)
01-May-2020
2.183
(1.988,2.378)
01-Jun-2020
2.131
(1.948,2.314)
01-Jul-2020
2.101
(1.917,2.285)
01-Aug-2020
2.026
(1.846,2.206)
01-Sep-2020
1.994
(1.806,2.182)
01-Oct-2020
1.987
(1.792,2.182)
01-Nov-2020
2.02
(1.824,2.216)
01-Dec-2020
2.037
(1.833,2.24)
01-Jan-2021
2.007
(1.793,2.221)
01-Feb-2021
2.012
(1.797,2.228)
01-Mar-2021
2.005
(1.789,2.221)
01-Apr-2021
2
(1.785,2.215)
01-May-2021
2.004
(1.789,2.219)
01-Jun-2021
2.021
(1.806,2.236)


cases
Taylor's law parameters of cumulative U.S. COVID-19 cases/county by state
$s_n^2 = a\bar{X}_n^{\ b}$
Coefficient a
$\hat{a}, 95\% CI$
$\bar{a}$
intercept a
30
25
20
15
10
5
0
Apr 2020
Jul 2020
Oct 2020
Jan 2021
Apr 2021
Jul 2021
month
Exponent b
$\hat{b}, 95\% CI$
$\bar{b}$
$b = 2$
slope b
2.4
2.3
2.2
2.1
2
1.9
1.8
Apr 2020
Jul 2020
Oct 2020
Jan 2021
Apr 2021
Jul 2021
Apr 2020
month
Jul 2021

deaths



Apr 2020 Jul 2021

Taylor's law describes counties' cumulative cases & deaths.

From April 2020 onward, TL holds:

1. log variance of counts (over counties) increases linearly with the log mean of counts (over counties) from state to state.

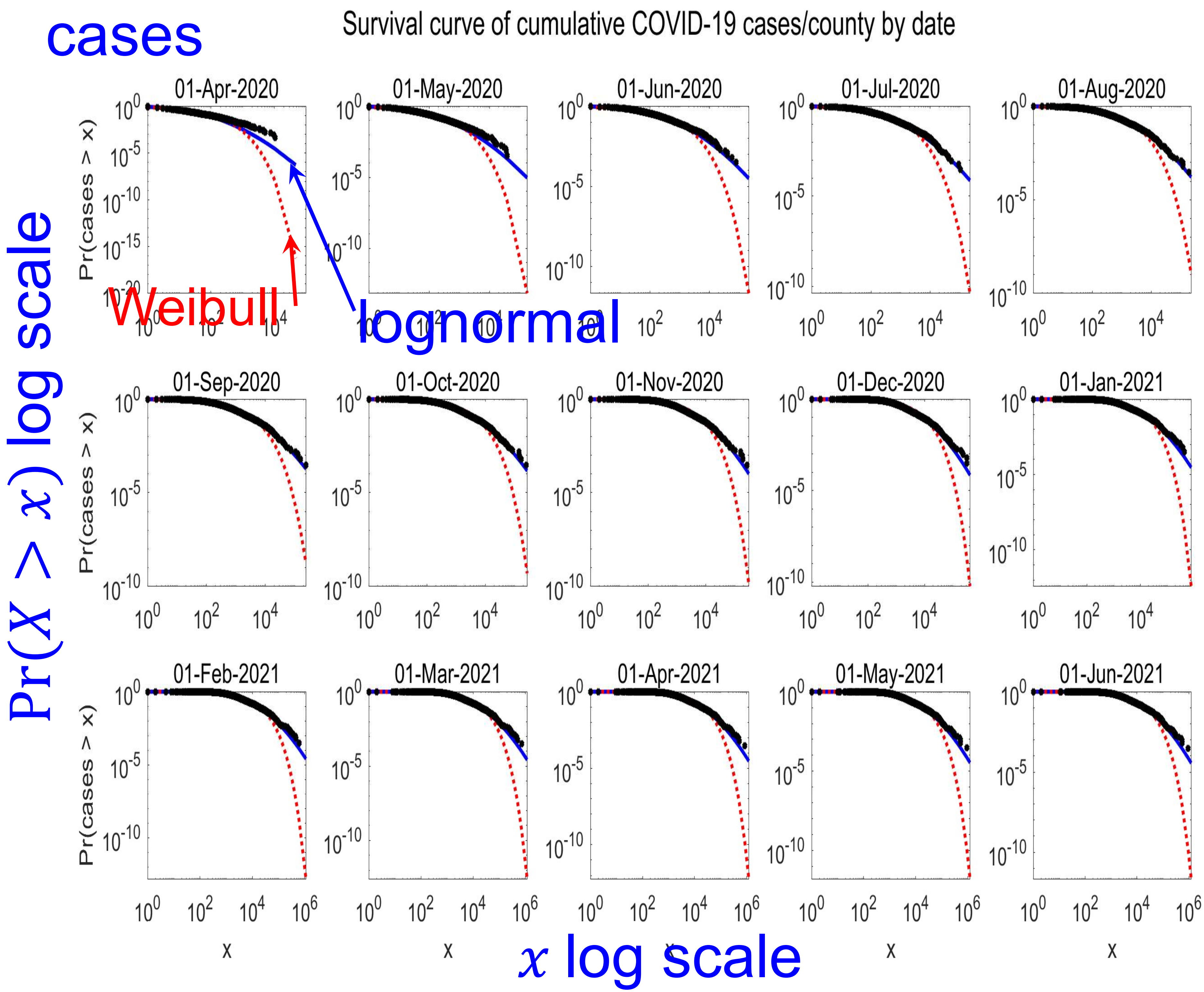
2. Slope $b \approx 2$.

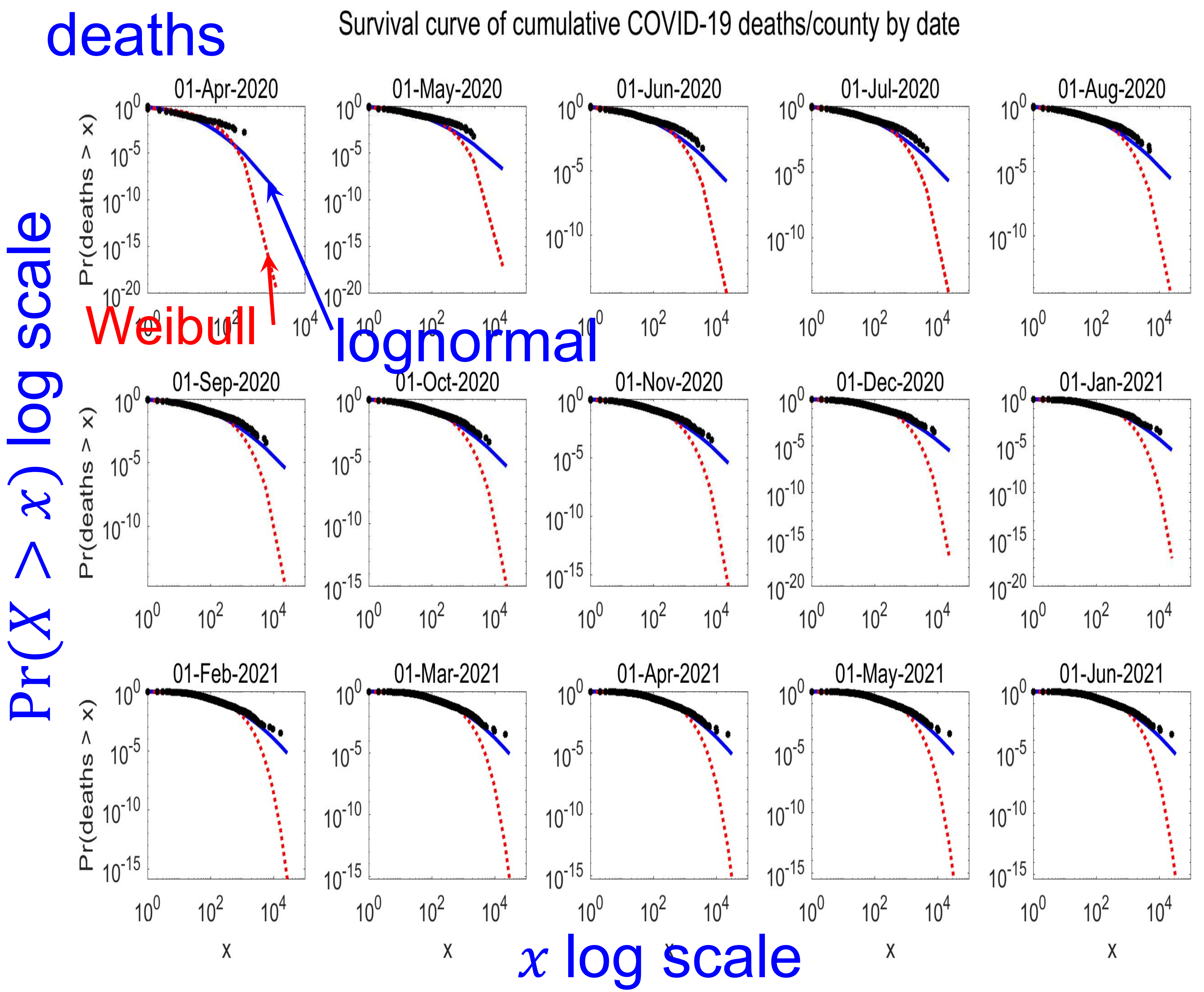
Why?

Survival curve plots probability that counts > x as a function of x.

To cover wide ranges of probability & of counts, we plot log(Pr{X > x}) as a function of log(x).

We also fit lognormal & Weibull distributions by maximum likelihood to counts of all counties.


cases
Survival curve of cumulative COVID-19 cases/county by date
01-Apr-2020
01-May-2020
01-Jun-2020
01-Jul-2020
01-Aug-2020
01-Sep-2020
01-Oct-2020
01-Nov-2020
01-Dec-2020
01-Jan-2021
01-Feb-2021
01-Mar-2021
01-Apr-2021
01-May-2021
01-Jun-2021
Pr(cases > x)
x
Weibull
lognormal
$\Pr(X > x)$ log scale
x log scale



deaths
Survival curve of cumulative COVID-19 deaths/county by date
01-Apr-2020
01-May-2020
01-Jun-2020
01-Jul-2020
01-Aug-2020
01-Sep-2020
01-Oct-2020
01-Nov-2020
01-Dec-2020
01-Jan-2021
01-Feb-2021
01-Mar-2021
01-Apr-2021
01-May-2021
01-Jun-2021
Pr(deaths > x)
x
Weibull
lognormal
$\Pr(X > x)$ log scale
x log scale

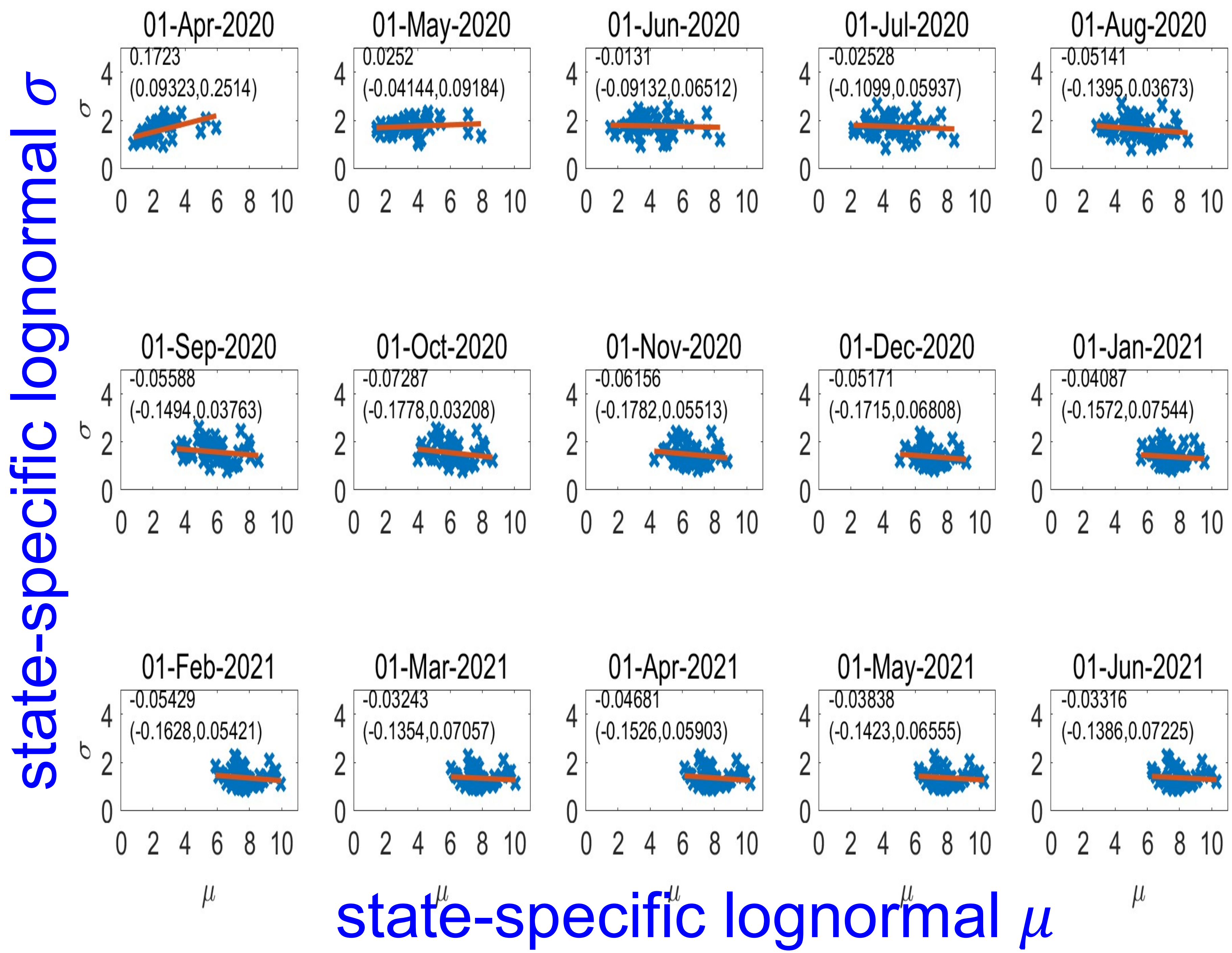
Lognormal describes >99% of distributions of cases & deaths.

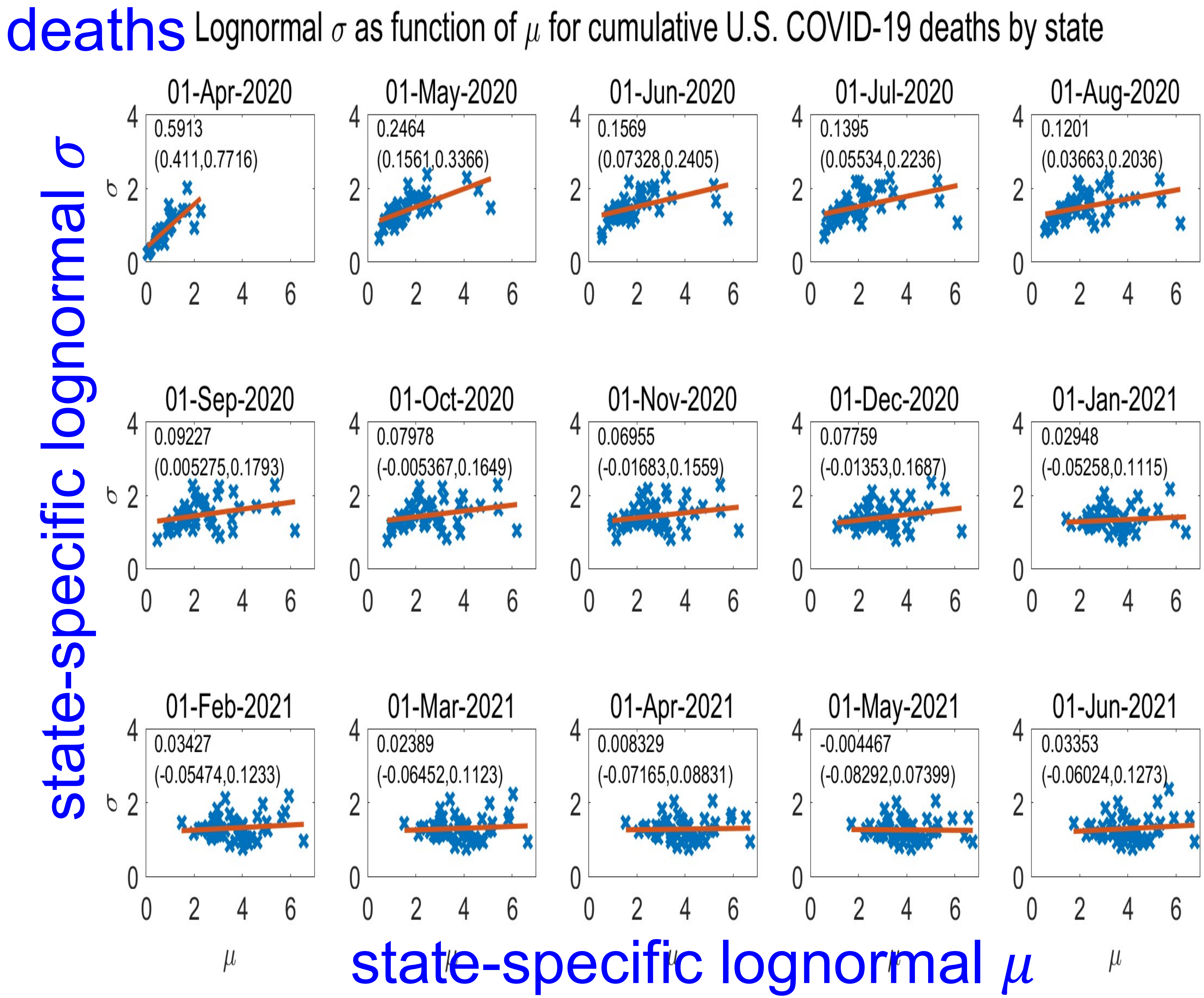
Many models lead to lognormal & Weibull distributions.

Lognormal is much closer than Weibull to data's upper tail, but also falls too fast.

If count has lognormal distribution, if σ^2 is constant, & if only μ varies from state to state, then Taylor's law holds with slope 2.

cases Lognormal σ as function of μ for cumulative U.S. COVID-19 cases by state



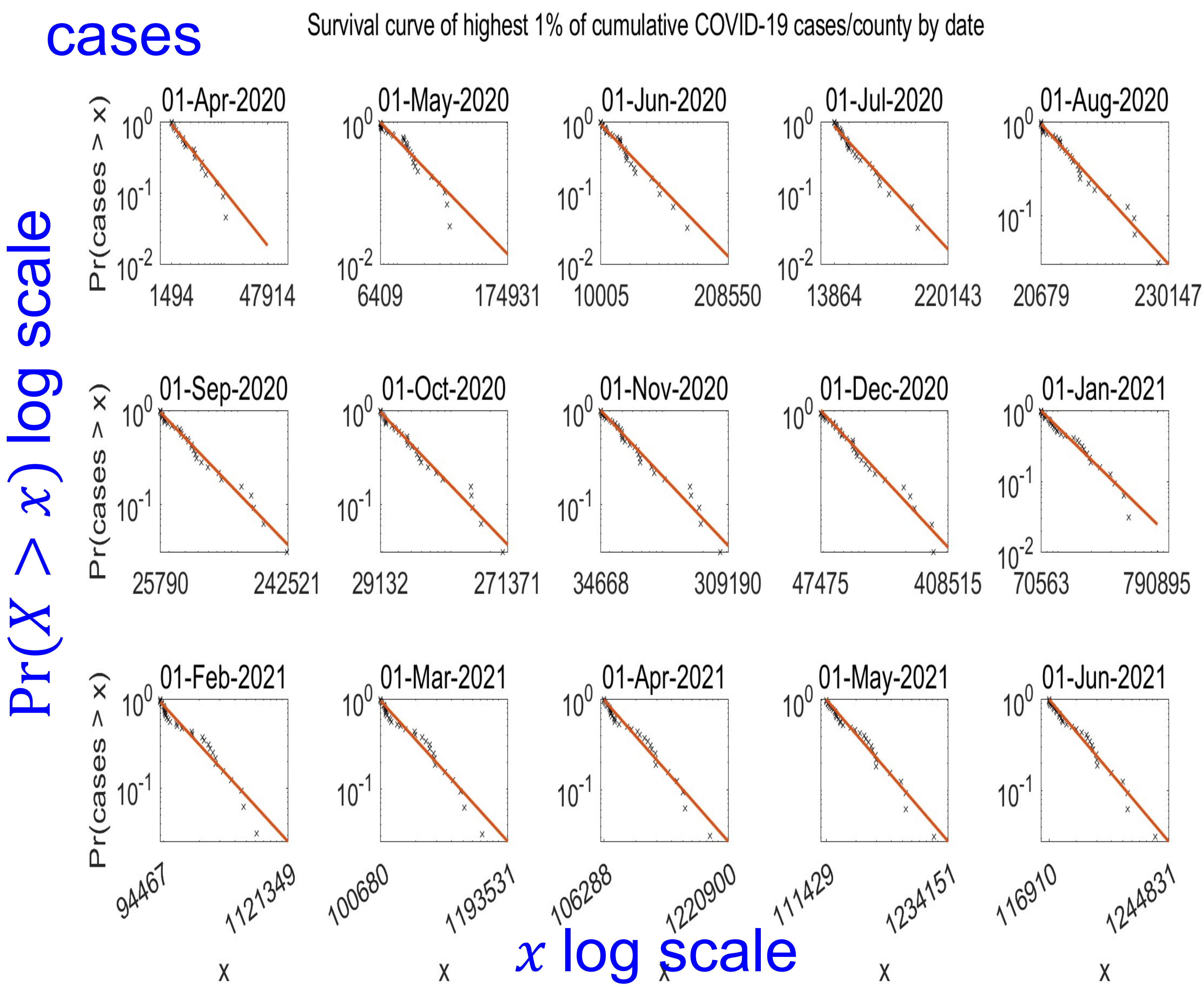

deaths
Lognormal σ as function of μ for cumulative U.S. COVID-19 deaths by state
01-Apr-2020
0.5913
(0.411,0.7716)
01-May-2020
0.2464
(0.1561,0.3366)
01-Jun-2020
0.1569
(0.07328,0.2405)
01-Jul-2020
0.1395
(0.05534,0.2236)
01-Aug-2020
0.1201
(0.03663,0.2036)
01-Sep-2020
0.09227
(0.005275,0.1793)
01-Oct-2020
0.07978
(-0.005367,0.1649)
01-Nov-2020
0.06955
(-0.01683,0.1559)
01-Dec-2020
0.07759
(-0.01353,0.1687)
01-Jan-2021
0.02948
(-0.05258,0.1115)
01-Feb-2021
0.03427
(-0.05474,0.1233)
01-Mar-2021
0.02389
(-0.06452,0.1123)
01-Apr-2021
0.008329
(-0.07165,0.08831)
01-May-2021
-0.004467
(-0.08292,0.07399)
01-Jun-2021
0.03353
(-0.06024,0.1273)
σ
μ
state-specific lognormal σ
state-specific lognormal μ

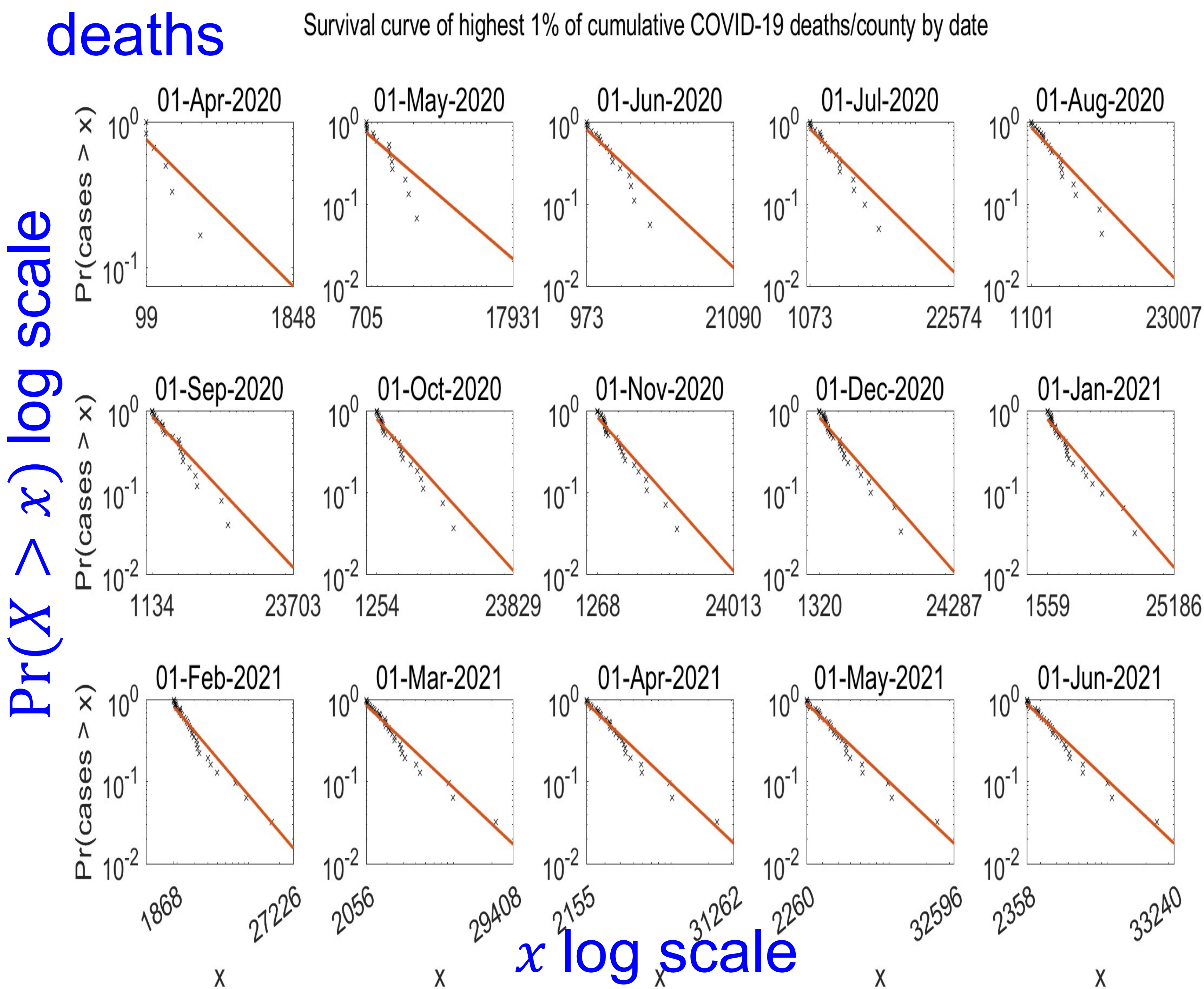
Lognormal explains TL with slope 2 for lower 99% of counts but not 1% upper tail.

Lognormal μ varies much more widely than lognormal σ^2, generating predicted means & predicted variances that closely approximate Taylor's law with slope 2.

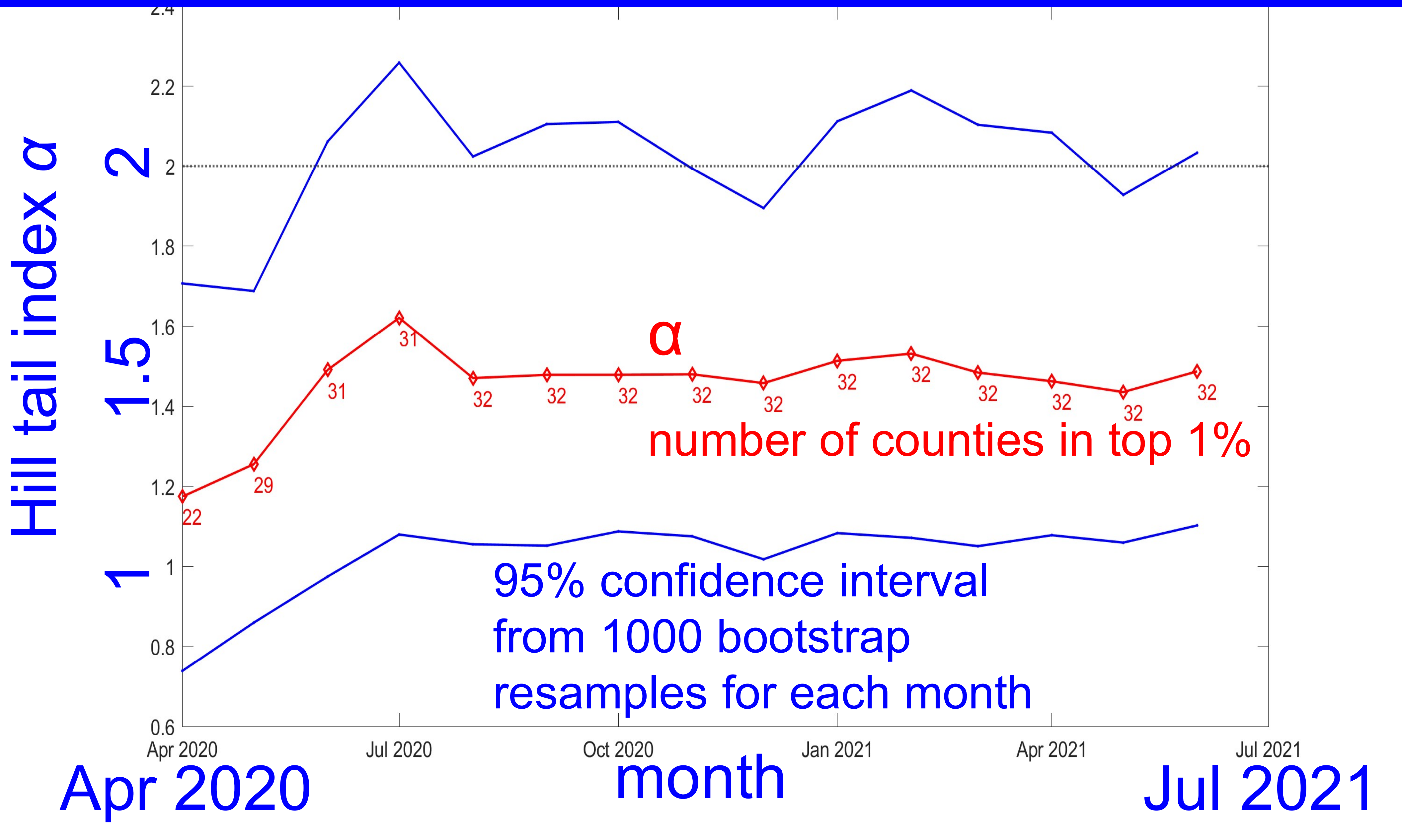
But the very largest counts of cases & deaths are more extreme than lognormal distribution predicts.

We zoom in to the counties with the highest 1% of counts of cases or deaths.

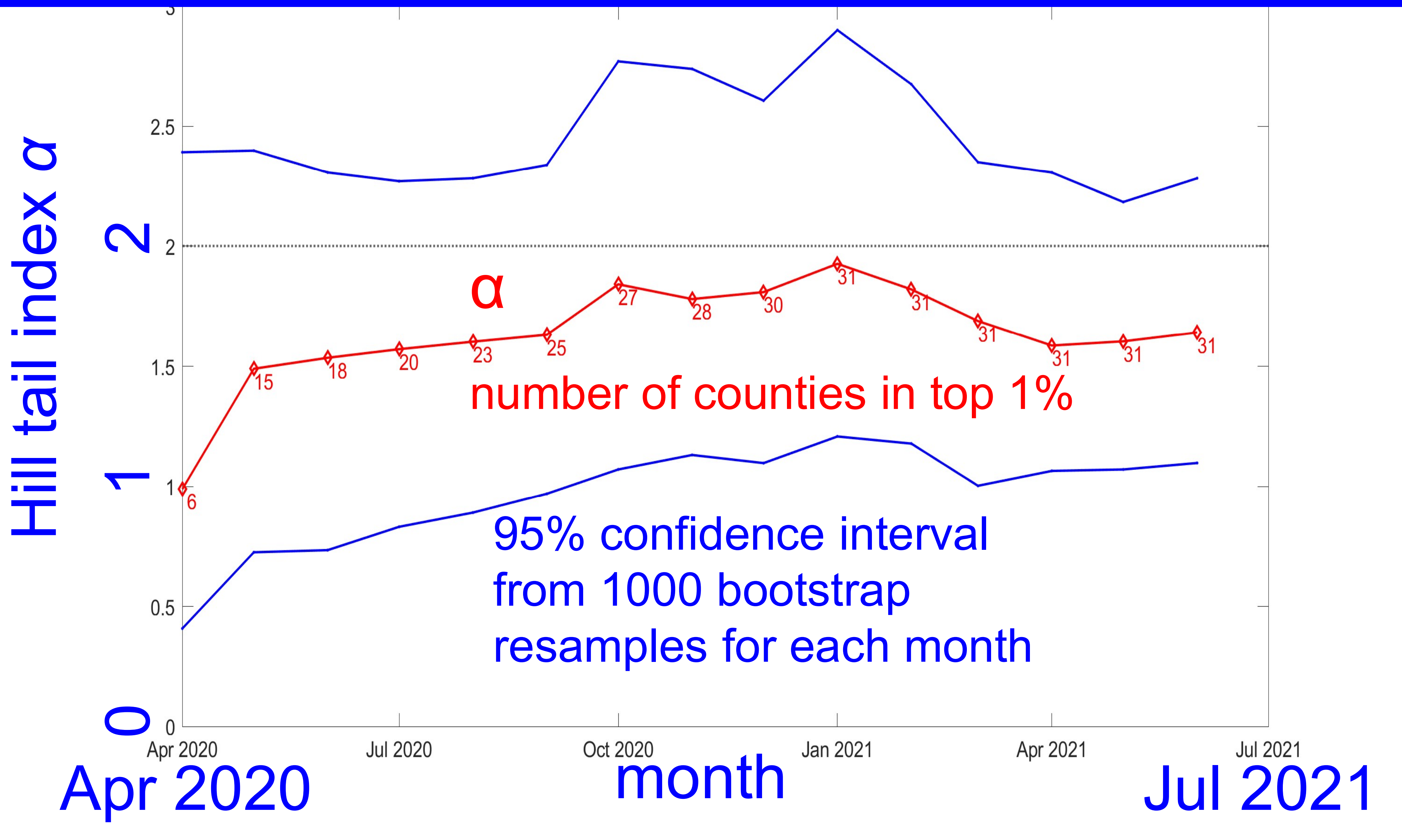

cases
Survival curve of highest 1% of cumulative COVID-19 cases/county by date
01-Apr-2020
01-May-2020
01-Jun-2020
01-Jul-2020
01-Aug-2020
01-Sep-2020
01-Oct-2020
01-Nov-2020
01-Dec-2020
01-Jan-2021
01-Feb-2021
01-Mar-2021
01-Apr-2021
01-May-2021
01-Jun-2021
Pr(cases > x)
1494 47914
6409 174931
10005 208550
13864 220143
20679 230147
25790 242521
29132 271371
34668 309190
47475 408515
70563 790895
94467 1121349
100680 1193531
106288 1220900
111429 1234151
116910 1244831
x
$\Pr(X > x)$ log scale
x log scale


deaths
Survival curve of highest 1% of cumulative COVID-19 deaths/county by date
01-Apr-2020
01-May-2020
01-Jun-2020
01-Jul-2020
01-Aug-2020
01-Sep-2020
01-Oct-2020
01-Nov-2020
01-Dec-2020
01-Jan-2021
01-Feb-2021
01-Mar-2021
01-Apr-2021
01-May-2021
01-Jun-2021
Pr(cases > x)
x
Pr(X > x) log scale
x log scale

For counties with highest 1% of cases, Hill estimates of tail index: $1<\alpha<2$



For counties with highest 1% of deaths, Hill estimates of tail index: 1<α<2



Empirical survival curves suggest variance is infinite.

The estimated upper tail index is $1 < \alpha < 2$ in all 15 months for cases & all but first month April 2020 for deaths, so

variance is infinite,

mean is finite.

"Wonder / Fear / Astonishment"

Regularly varying upper tail with index $\alpha \in (1,2)$ explains why TL with $b = 2$ holds even for largest counts where lognormal distribution fails.

Simulations: 100 samples of size 100

4 models in RV(α): $|N(0,1)|^{-\alpha^{-1}}$, $|U|^{-\alpha^{-1}}$, $|U_1 U_2|^{-\alpha^{-1}}$, $|U_1 U_2 U_3|^{-\alpha^{-1}}$

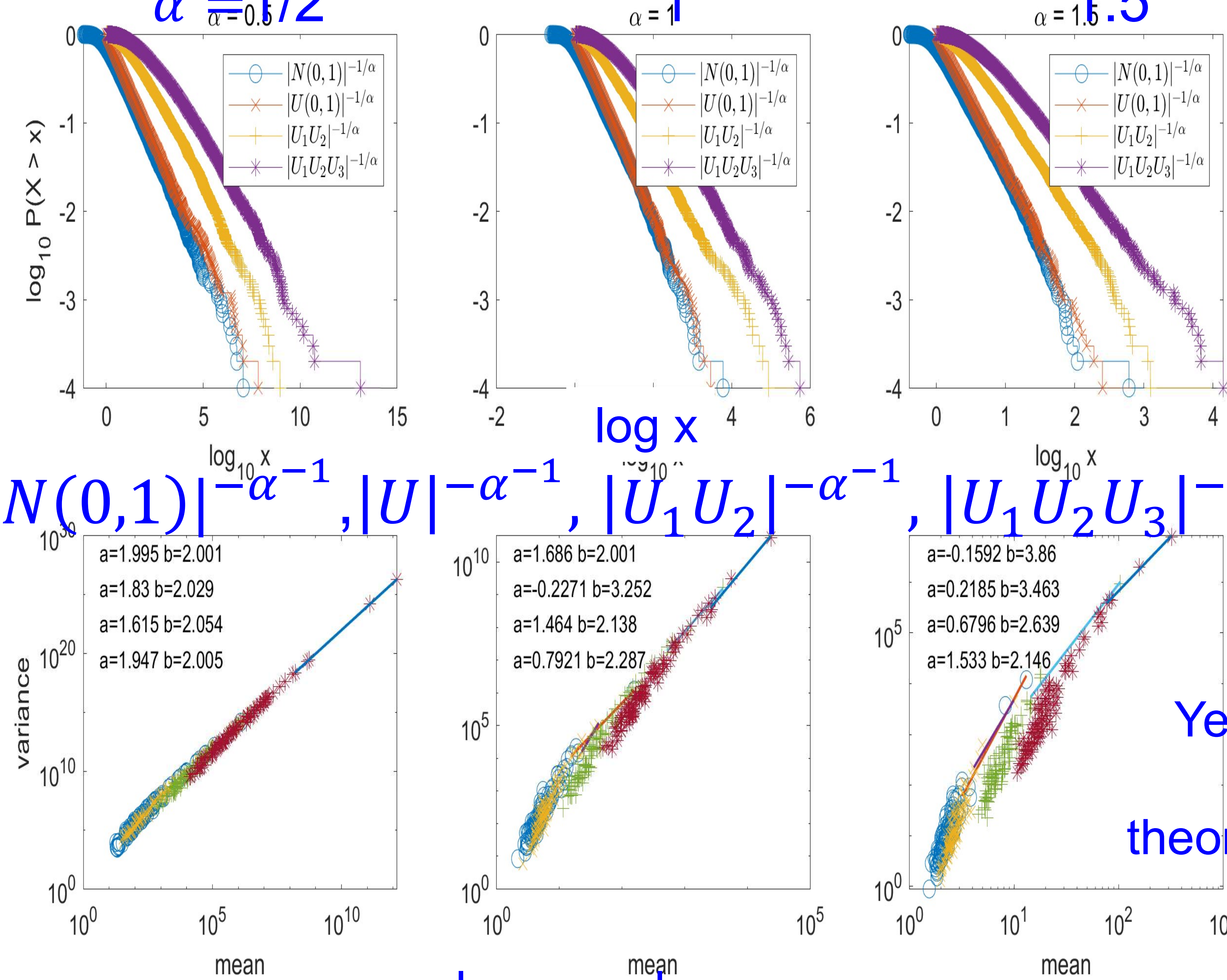
$\alpha =$1/2, 1, 3/2, with & without dependence.

TL with $b \to 2$ is confirmed by mathematics.

4 models in RV(α), 100 samples of size 100

$\alpha = 1/2$ 1   1.5

log Pr(X>x)



log x

$|N(0,1)|^{-\alpha^{-1}}, |U|^{-\alpha^{-1}}, |U_1U_2|^{-\alpha^{-1}}, |U_1U_2U_3|^{-\alpha^{-1}}$

log sample variance

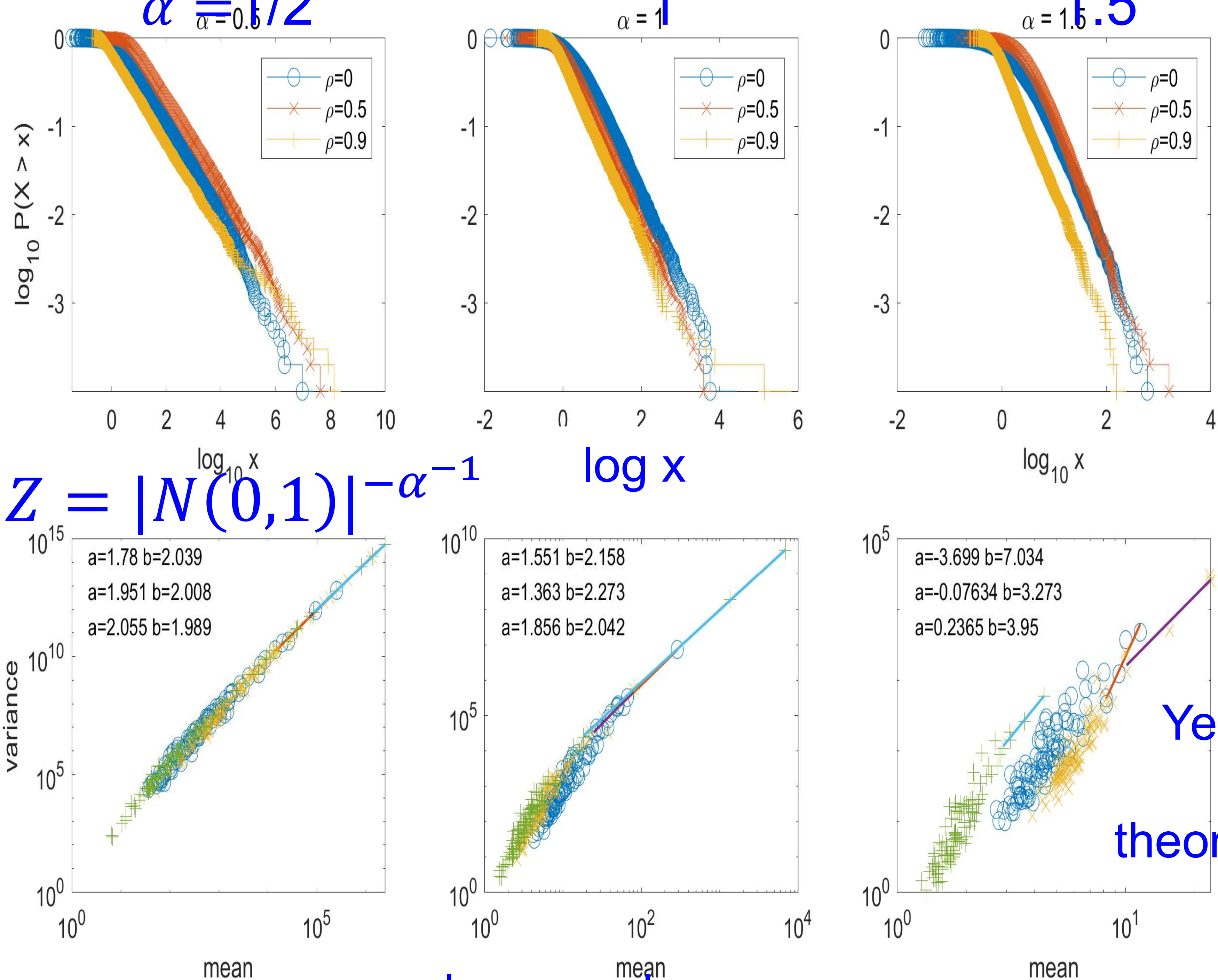
Yes, we have theorems.

log sample mean

ρ=0, 0.5, 0.9, 100 samples of size 100

α = 1/2   1   1.5

log Pr(X>x)



log x

$$Z = |N(0,1)|^{-\alpha^{-1}}$$

log sample variance

log sample mean

Yes, we have theorems.

So what?

If the variances of cases & deaths per county are infinite, facility & resource planning should prepare for unboundedly high counts.

No single county (or state, or other jurisdiction) can prepare for unboundedly high counts.

Cooperative exchanges of support should be planned cooperatively.

My math collaborators & teachers

Thank you!
Questions?

cohen@rockefeller.edu

20190906 La Fage To Florac
Cévennes "Cham des Bondons
Chabusse"

2/3/2024