COVID-19 cases and deaths in the United States, Taylor's law of fluctuation scaling, and heavy tails

Joint work with
Richard A. Davis & Gennady Samorodnitsky

Joel E. Cohen, cohen@rockefeller.edu
Rockefeller University & Columbia University

2021-09-17, Ecodep Conference (Paul Doukhan),
Paris Seine University, Cergy-Pontoise, France

Ashokan Reservoir, Ulster County, NY, JEC 20151015

Outline

1. COVID-19 U.S. data analysis:
   Taylor's law & heavy upper tails
2. Simulations of an idealization
3. Mathematics

A global pandemic of COVID-19

By 12 September 2021, coronavirus SARS-CoV-2 caused >225 million reported cases of COVID-19 disease and >4.6 million deaths.

U.S. reported more cases (41.8 million) and more deaths (678,000) than any other country.

https://www.worldometers.info/coronavirus/countries-where-coronavirus-has-spread/

Administrative structure of U.S.

~56 first-level divisions := "states": states, territories, possessions, Washington DC

~3100 second-level subdivisions := "counties":

"… 3,006 counties; 14 boroughs and 11 census areas in Alaska; … 64 parishes in Louisiana; Baltimore city, Maryland; St. Louis city, Missouri; that part of Yellowstone National Park in Montana; Carson City, Nevada; and 41 independent cities in Virginia."

U.S. Census, *Geographic Areas Reference Manual*, 1994

~55 counties/state ≈ ~number of states

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

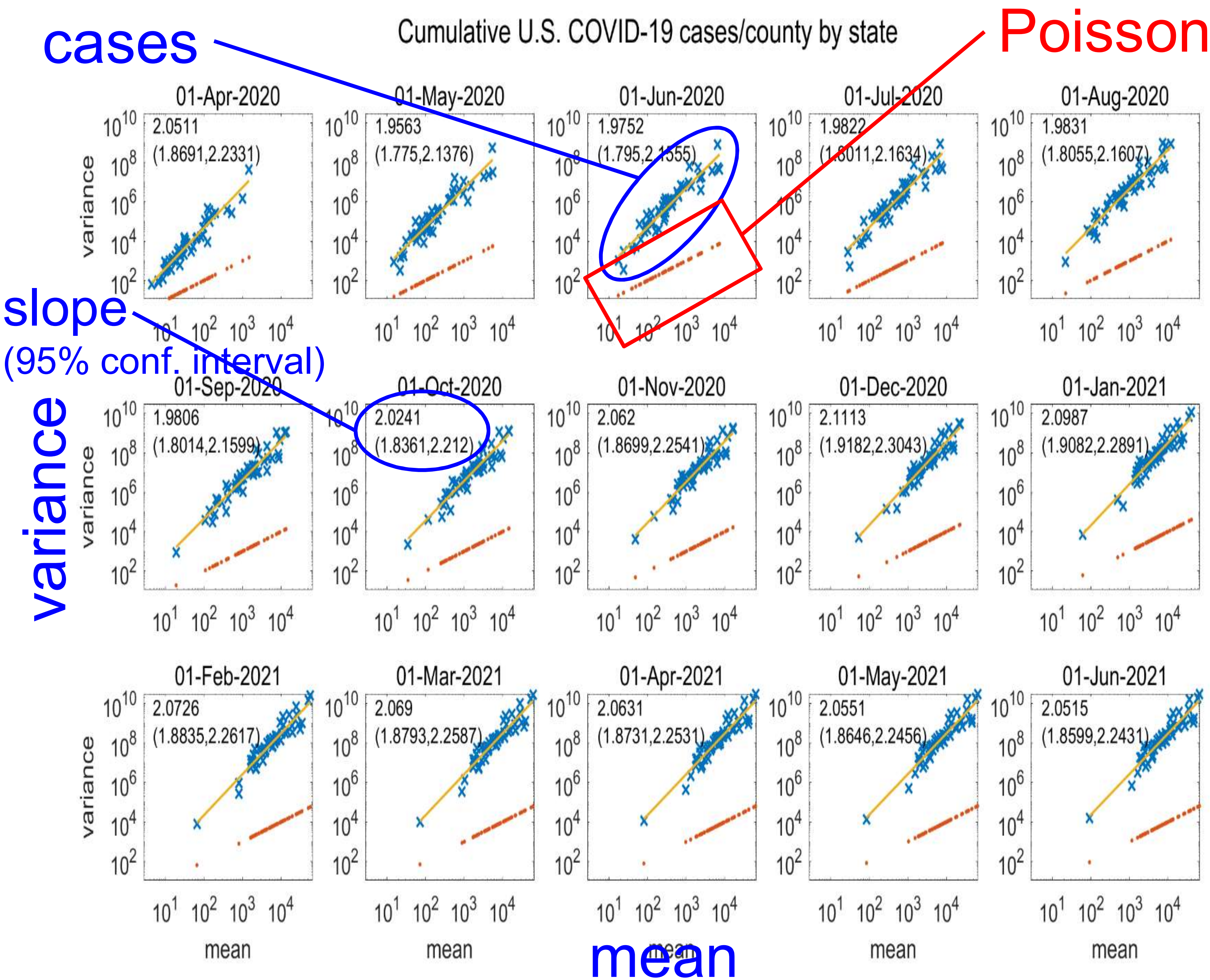
On each date, cumulative cases & deaths within each state by county

State number →	j=1	j=2	j=3	j=…
County 1	x_{11}	x_{12}	x_{13}	$x_{\ldots}$
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…

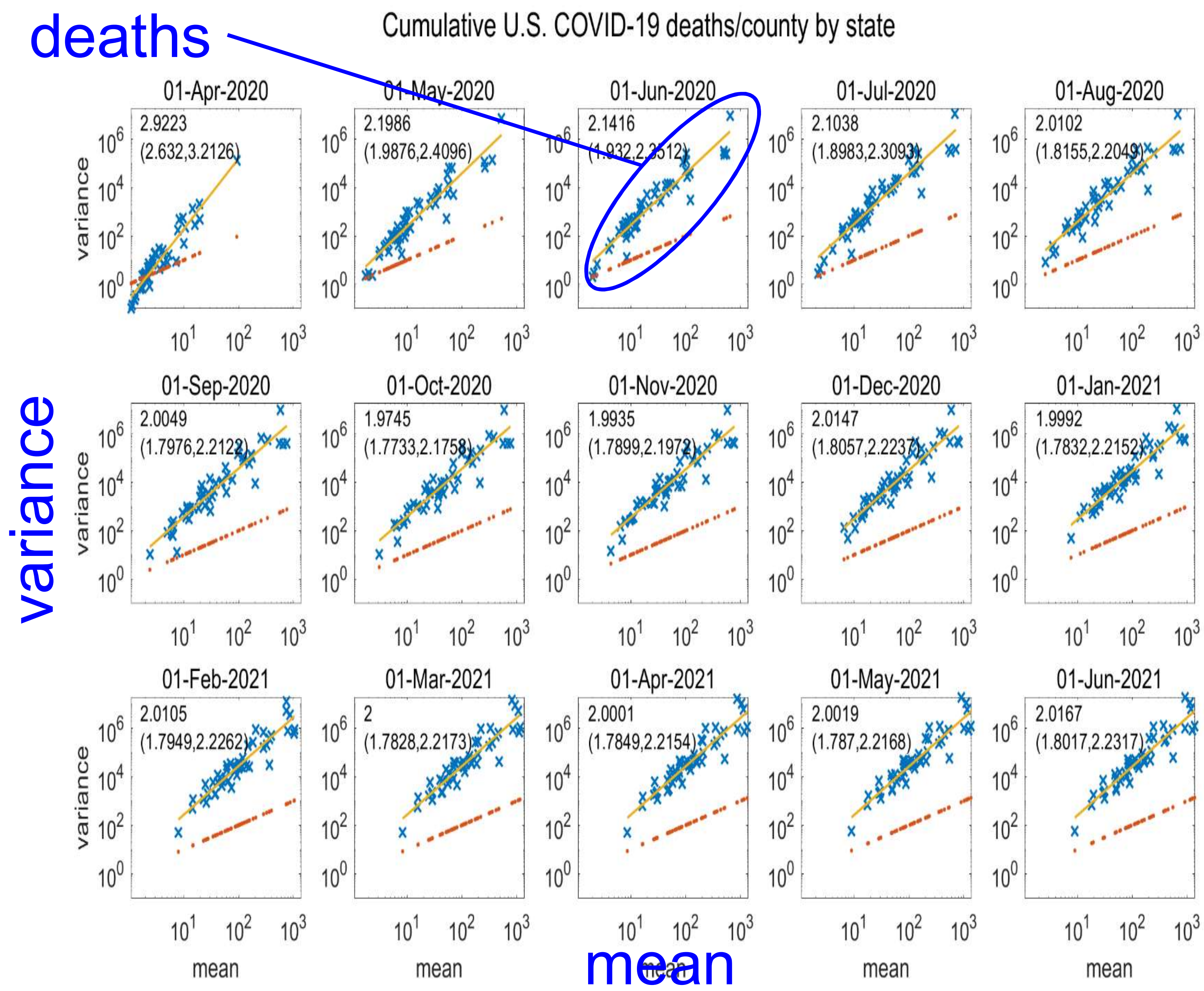
Cumulative U.S. cases/county by "state" as of June 1, 2021

(from lowest to highest mean)

state	number of cases	mean cases per county	variance cases per county
Virgin Islands	3465	1155	6.7595e+05
Vermont	24224	1614.9	3.1139e+06
South Dakota	1.2419e+05	1881.7	1.8494e+07
⋮	⋮	⋮	⋮
Massachusetts	7.0713e+05	47142	1.7181e+09
Arizona	8.8145e+05	58764	1.9367e+10
California	3.791e+06	65363	3.03e+10



Cumulative U.S. COVID-19 cases/county by state
cases
Poisson
slope
(95% conf. interval)
variance
mean
01-Apr-2020
2.0511
(1.8691,2.2331)
01-May-2020
1.9563
(1.775,2.1376)
01-Jun-2020
1.9752
(1.795,2.1555)
01-Jul-2020
1.9822
(1.8011,2.1634)
01-Aug-2020
1.9831
(1.8055,2.1607)
01-Sep-2020
1.9806
(1.8014,2.1599)
01-Oct-2020
2.0241
(1.8361,2.212)
01-Nov-2020
2.062
(1.8699,2.2541)
01-Dec-2020
2.1113
(1.9182,2.3043)
01-Jan-2021
2.0987
(1.9082,2.2891)
01-Feb-2021
2.0726
(1.8835,2.2617)
01-Mar-2021
2.069
(1.8793,2.2587)
01-Apr-2021
2.0631
(1.8731,2.2531)
01-May-2021
2.0551
(1.8646,2.2456)
01-Jun-2021
2.0515
(1.8599,2.2431)



deaths
variance
mean
Cumulative U.S. COVID-19 deaths/county by state
01-Apr-2020
2.9223
(2.632,3.2126)
01-May-2020
2.1986
(1.9876,2.4096)
01-Jun-2020
2.1416
(1.932,2.3512)
01-Jul-2020
2.1038
(1.8983,2.3093)
01-Aug-2020
2.0102
(1.8155,2.2049)
01-Sep-2020
2.0049
(1.7976,2.2122)
01-Oct-2020
1.9745
(1.7733,2.1758)
01-Nov-2020
1.9935
(1.7899,2.1972)
01-Dec-2020
2.0147
(1.8057,2.2237)
01-Jan-2021
1.9992
(1.7832,2.2152)
01-Feb-2021
2.0105
(1.7949,2.2262)
01-Mar-2021
2
(1.7828,2.2173)
01-Apr-2021
2.0001
(1.7849,2.2154)
01-May-2021
2.0019
(1.787,2.2168)
01-Jun-2021
2.0167
(1.8017,2.2317)

Major findings: counties' cumulative counts of cases or deaths

After first few months,

log variance (over counties) increases, state by state, linearly as a function of log mean (over counties), and

slope is close to (not significantly different from) 2.

Why?

TL data structure: multiple samples, each with multiple observations

Sample number →	j=1	j=2	j=3	j=…
Counts or nonnegative quantities	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 = average	m_1	m_2	m_3	m…
Variance	v_1	v_2	v_3	v…

Taylor's law *Nature* 1961

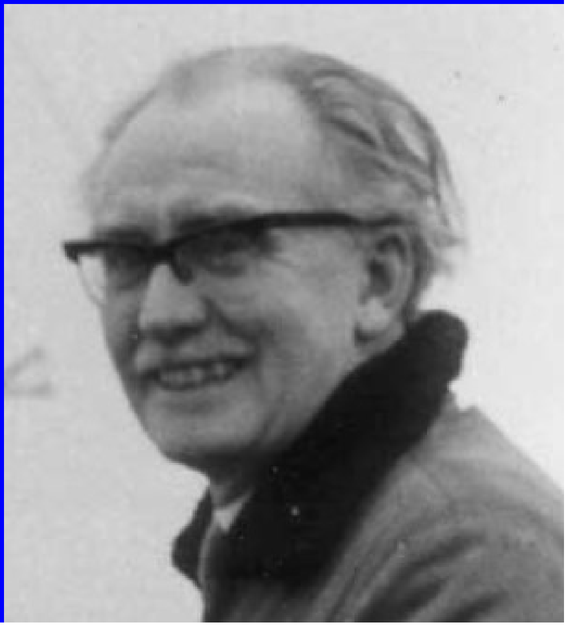
In multiple samples, the variance is proportional to a power of the mean.

$\text{variance} \approx a(\text{mean})^b, a > 0.$

$\log(\text{variance}) \approx \log(a) + b \cdot \log(\text{mean}).$

$\text{variance}/(\text{mean})^b \approx a, \quad a > 0.$

Taylor measured population density. The pattern applies more widely, but not universally. Taylor stated no model of deviations from exact equality.



Lionel Roy Taylor
(1924–2007)

2021-09-15

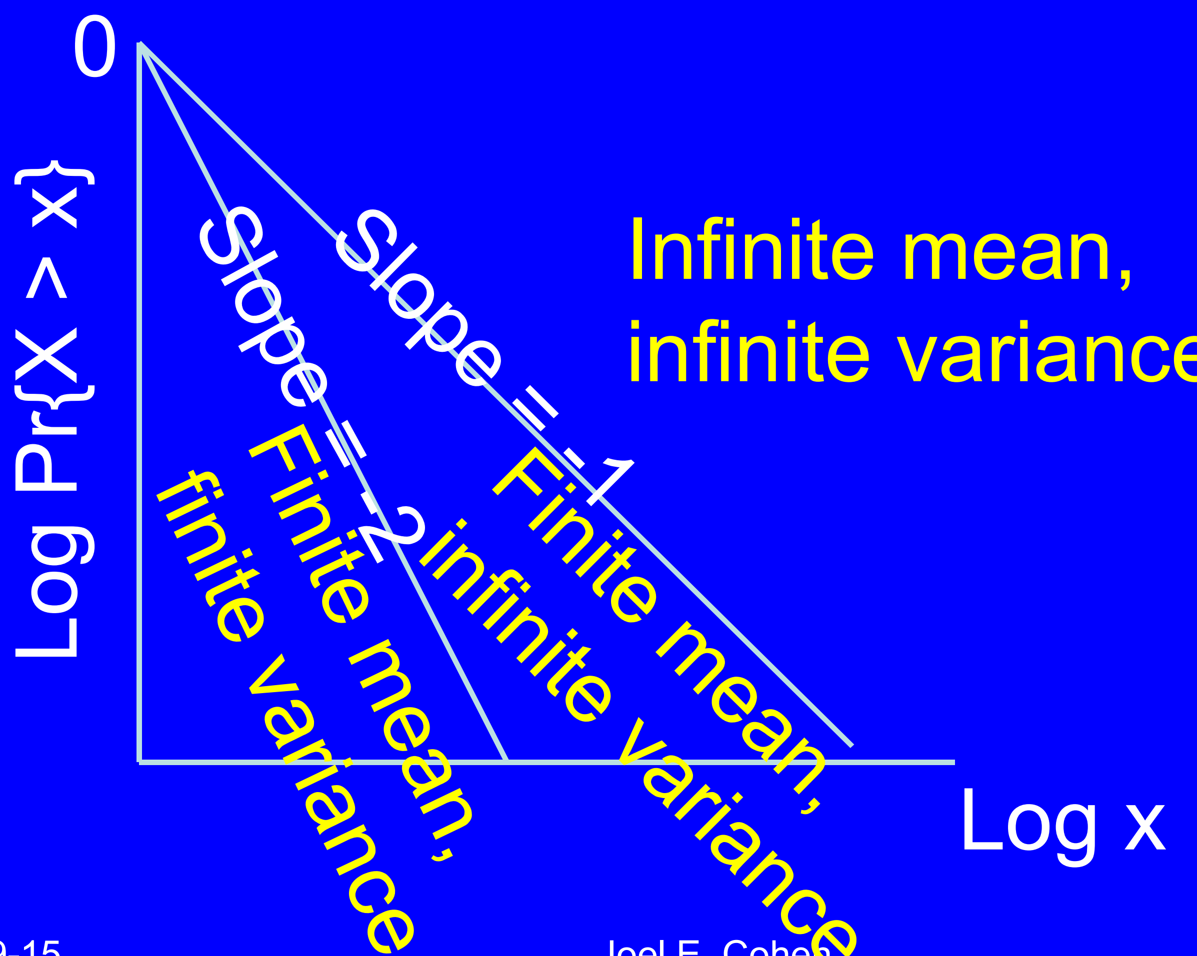
Heavy tails

Sample mean, sample variance are always finite. **But** some distributions have infinite mean or infinite variance.

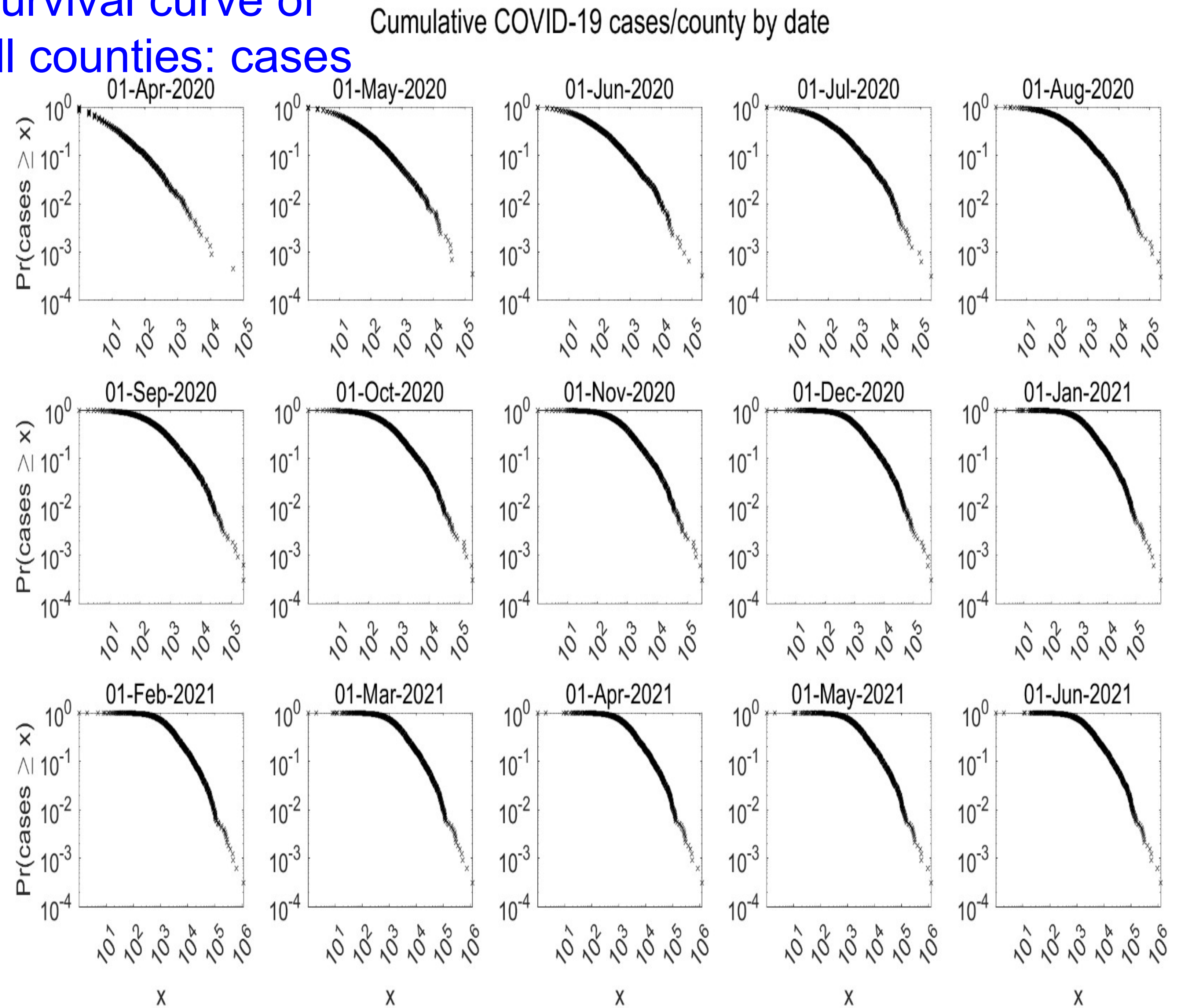
"Wonder / Fear / Astonishment"

Then sample mean (or sample variance) does not converge to a finite value, but instead moves to infinity, with increasing sample size.

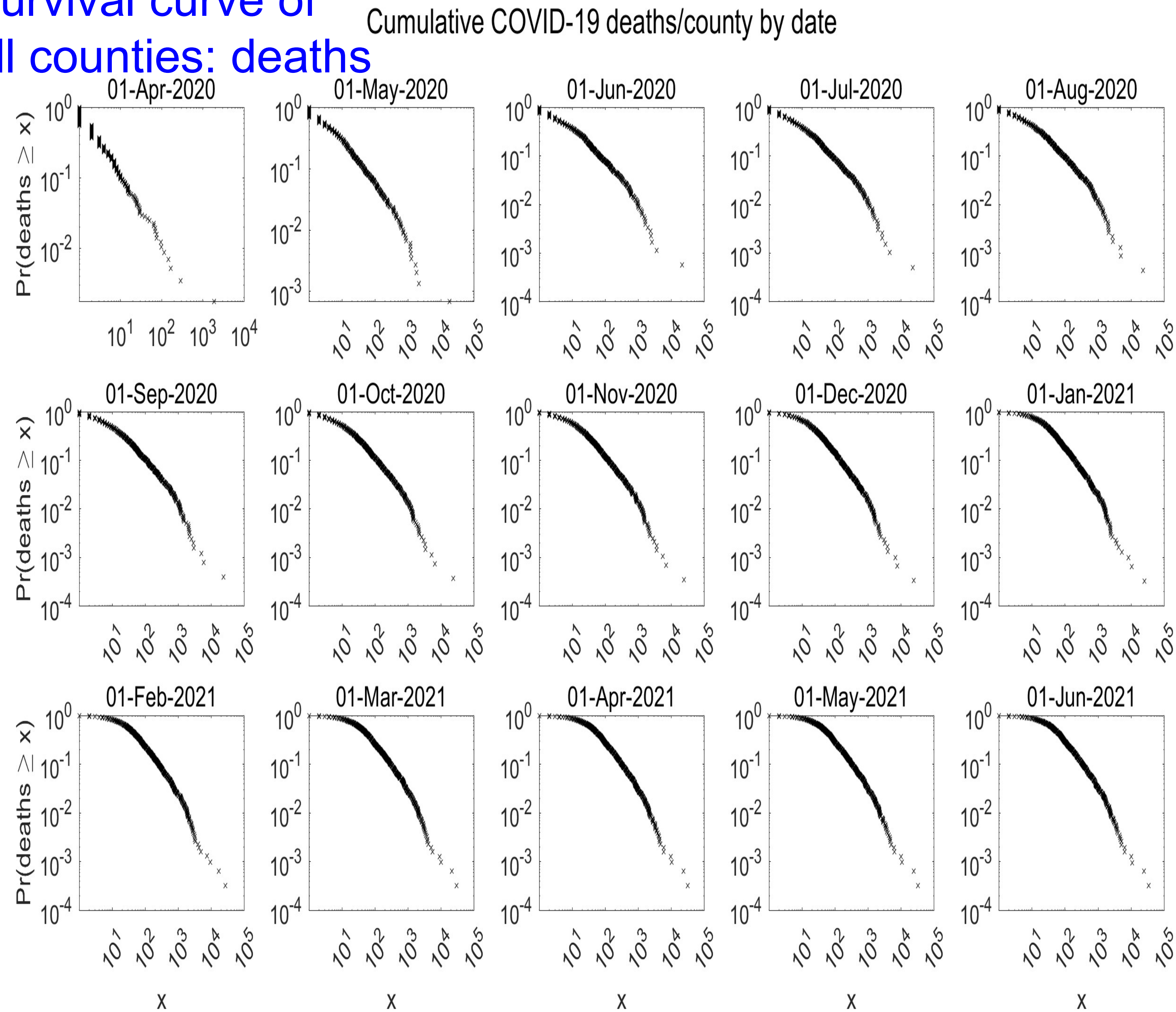
Upper tail of survival curve reveals which moments (if any) are finite.



Survival curve of all counties: cases

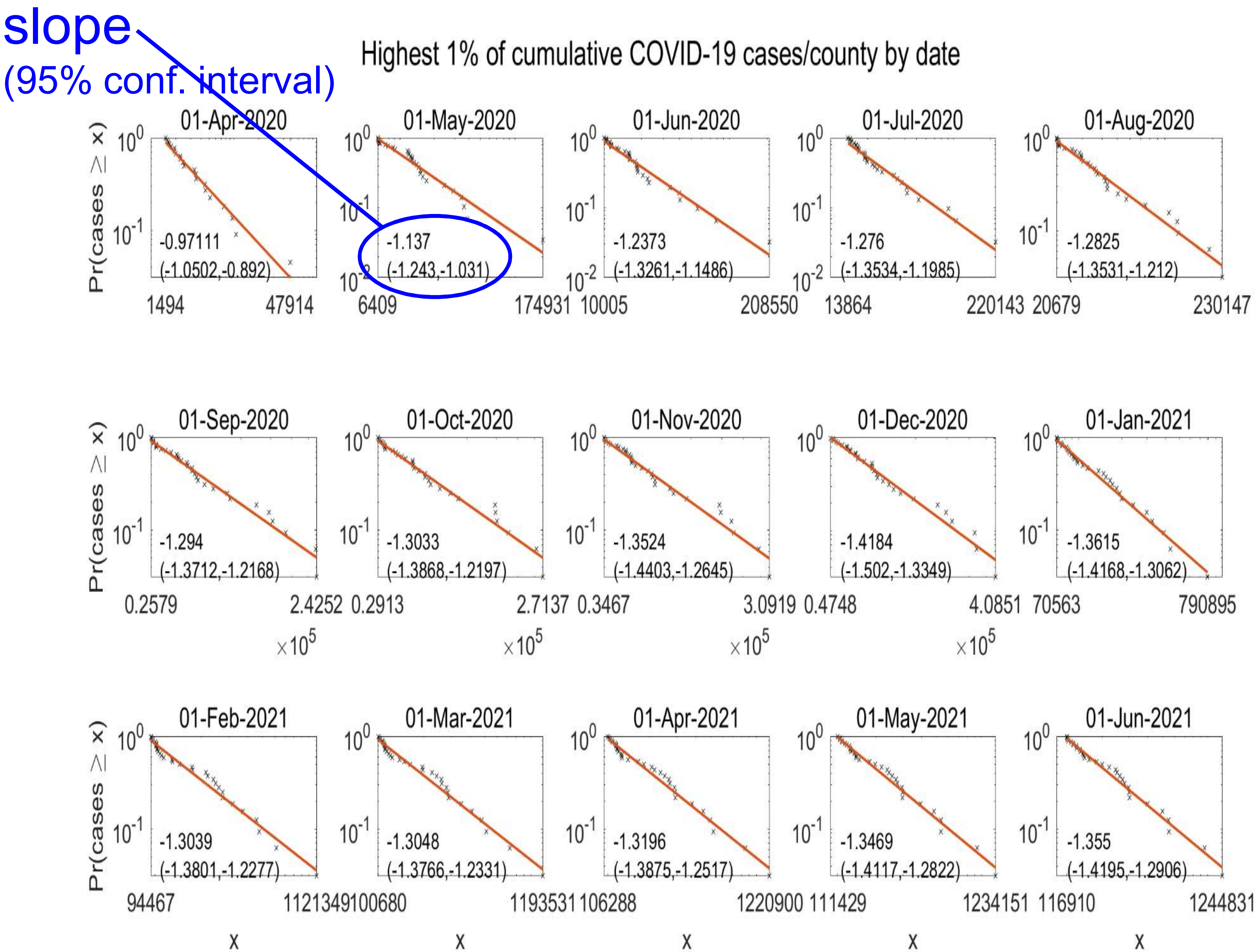


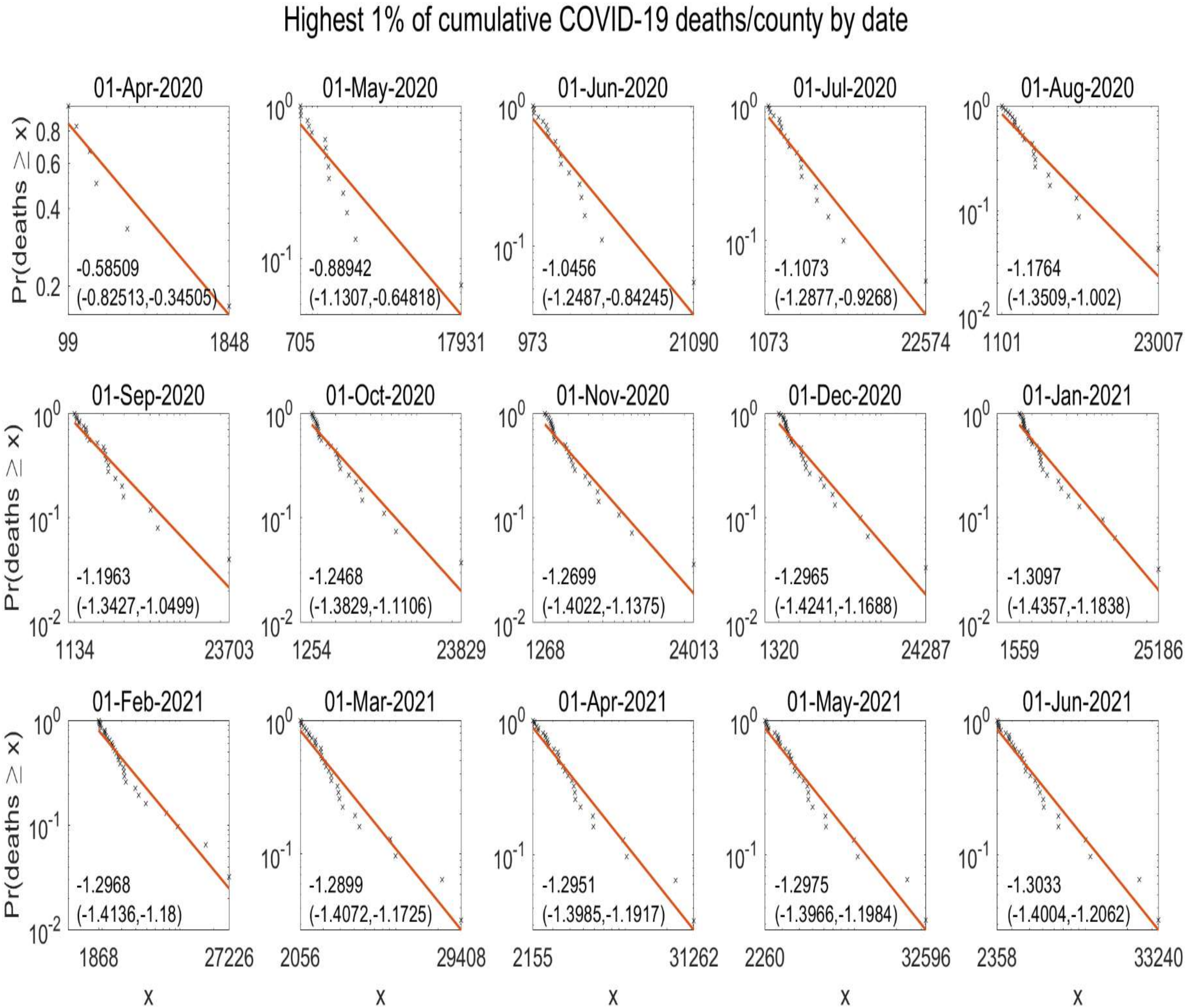
Survival curve of all counties: deaths



Cases and deaths by county are not Pareto-distributed over their entire range.

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


slope
(95% conf. interval)
Highest 1% of cumulative COVID-19 cases/county by date
Pr(cases ≥ x)
x
01-Apr-2020
-0.97111
(-1.0502,-0.892)
1494
47914
01-May-2020
-1.137
(-1.243,-1.031)
6409
174931
01-Jun-2020
-1.2373
(-1.3261,-1.1486)
10005
208550
01-Jul-2020
-1.276
(-1.3534,-1.1985)
13864
220143
01-Aug-2020
-1.2825
(-1.3531,-1.212)
20679
230147
01-Sep-2020
-1.294
(-1.3712,-1.2168)
0.2579
2.4252
01-Oct-2020
-1.3033
(-1.3868,-1.2197)
0.2913
2.7137
01-Nov-2020
-1.3524
(-1.4403,-1.2645)
0.3467
3.0919
01-Dec-2020
-1.4184
(-1.502,-1.3349)
0.4748
4.0851
01-Jan-2021
-1.3615
(-1.4168,-1.3062)
70563
790895
01-Feb-2021
-1.3039
(-1.3801,-1.2277)
94467
1121349
01-Mar-2021
-1.3048
(-1.3766,-1.2331)
100680
1193531
01-Apr-2021
-1.3196
(-1.3875,-1.2517)
106288
1220900
01-May-2021
-1.3469
(-1.4117,-1.2822)
111429
1234151
01-Jun-2021
-1.355
(-1.4195,-1.2906)
116910
1244831


Highest 1% of cumulative COVID-19 deaths/county by date
01-Apr-2020
-0.58509
(-0.82513,-0.34505)
99
1848
01-May-2020
-0.88942
(-1.1307,-0.64818)
705
17931
01-Jun-2020
-1.0456
(-1.2487,-0.84245)
973
21090
01-Jul-2020
-1.1073
(-1.2877,-0.9268)
1073
22574
01-Aug-2020
-1.1764
(-1.3509,-1.002)
1101
23007
01-Sep-2020
-1.1963
(-1.3427,-1.0499)
1134
23703
01-Oct-2020
-1.2468
(-1.3829,-1.1106)
1254
23829
01-Nov-2020
-1.2699
(-1.4022,-1.1375)
1268
24013
01-Dec-2020
-1.2965
(-1.4241,-1.1688)
1320
24287
01-Jan-2021
-1.3097
(-1.4357,-1.1838)
1559
25186
01-Feb-2021
-1.2968
(-1.4136,-1.18)
1868
27226
01-Mar-2021
-1.2899
(-1.4072,-1.1725)
2056
29408
01-Apr-2021
-1.2951
(-1.3985,-1.1917)
2155
31262
01-May-2021
-1.2975
(-1.3966,-1.1984)
2260
32596
01-Jun-2021
-1.3033
(-1.4004,-1.2062)
2358
33240
Pr(deaths ≥ x)
x

Major findings: highest 1% of counties' cumulative counts of cases or deaths

After first few months, in the highest 1% of counts of cases or deaths by county,

$\log \Pr\{count > x\}$ decreases linearly as $\log x$ increases, and

-1 > slope > -2.

Empirical survival curves suggest variance is infinite.

The slope of the upper tail on log-log coordinates is statistically significantly greater than -2 in all 15 months for cases & deaths. Variance is infinite in all cases.

The slope is statistically significantly less than -1 in 14 most recent months for cases & 11 most recent months for deaths. Except at beginning of epidemic, mean is finite.

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.

Wanted: theory of TL for multiple samples of heavy-tailed data

How do sample mean & sample variance behave in multiple samples with fixed or similar sample size, when:

data come from distributions with infinite variance or infinite mean, and

number of samples ≈ each sample size?

Outline

1. COVID-19 U.S. data analysis:
   Taylor's law & heavy upper tails
2. Simulations of an idealization
3. Mathematics

Probability distribution with regularly varying tail $\Pr(X > x)$

X is a regularly varying (RV) rv with index $\alpha > 0$ iff

$$\forall t > 0, \lim_{x \to \infty} \frac{\Pr(X > tx)}{\Pr(X > x)} = \frac{1}{t^{\alpha}}.$$

We write: $X \in RV(\alpha)$.

Then

$$\Pr(X > x) = \frac{L(x)}{x^{\alpha}}, \lim_{t \to \infty} \frac{\mathrm{L}(tx)}{L(t)} = 1, \forall x > 0.$$

Probability distribution with regularly varying tail $\Pr(X > x)$

Suppose $X \in RV(\alpha)$.

If $\alpha \in (0,1)$, then mean & all higher moments of X are infinite.

If $\alpha \in (1,2)$, then mean is finite, but variance & all higher moments of X are infinite.

If $\alpha \geq 2$, then X has finite mean & variance.

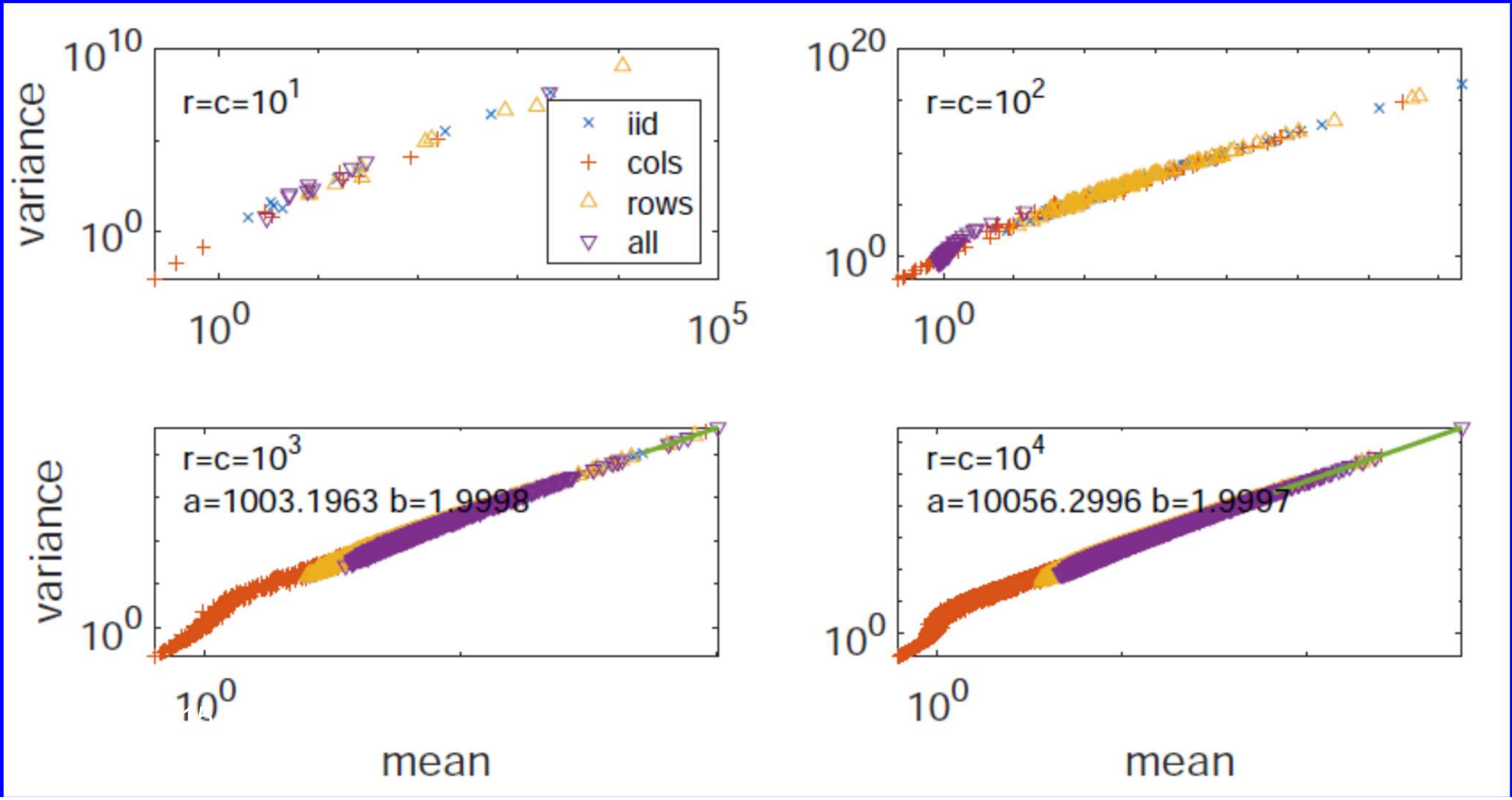
Simulations of idealized cases

We create $r \times c$ matrices, $r = c = 10, 100, 1000, 10^4, 10^5, 10^6$, with $RV(\alpha)$ elements (iid or asymptotically independent within columns, rows, or both), $\alpha \in (0,1) \cup (1,2)$.

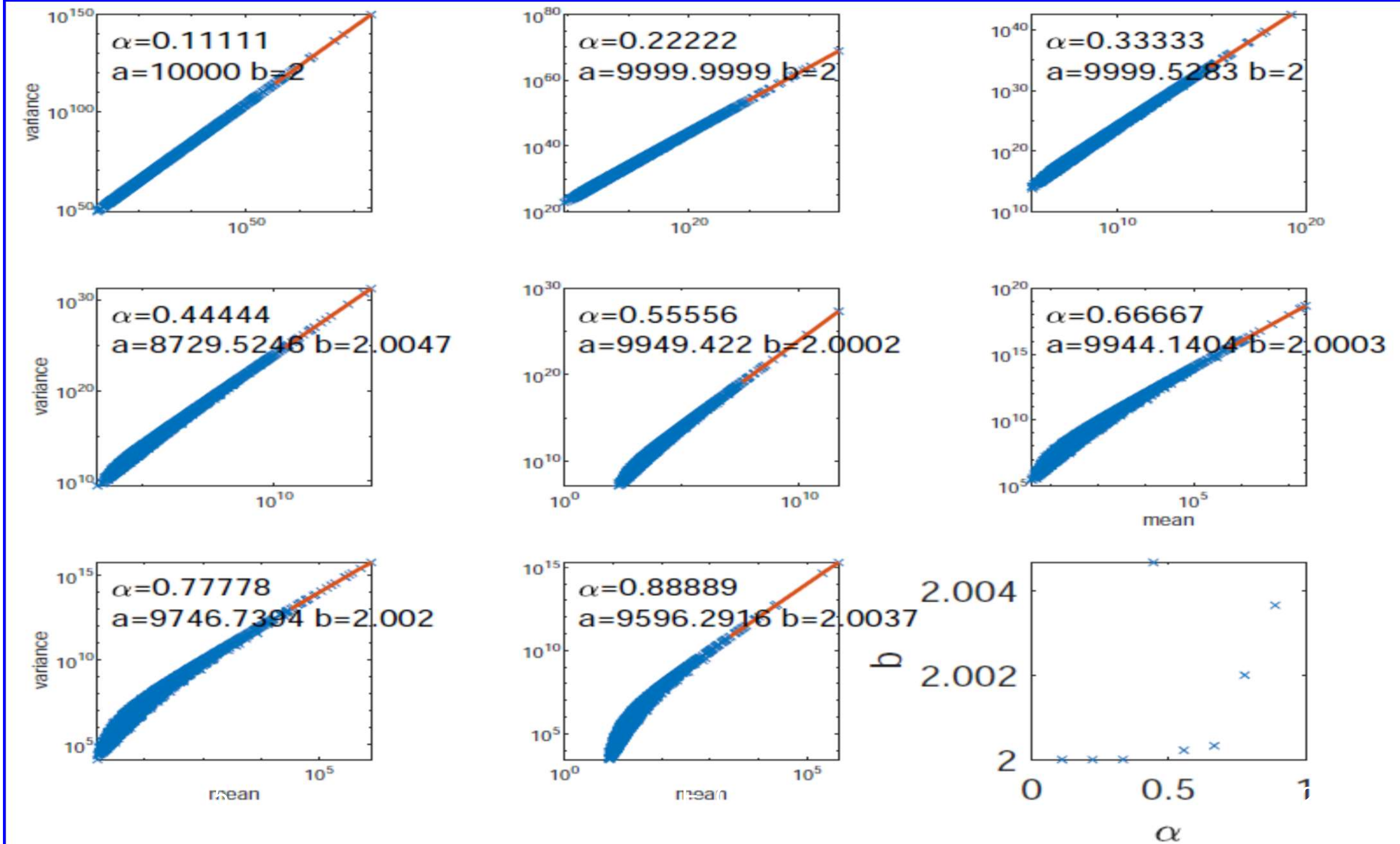
For each matrix, we plot log variance of each column as a function of log mean of the same column.

For larger means of large samples, slopes are close to 2.

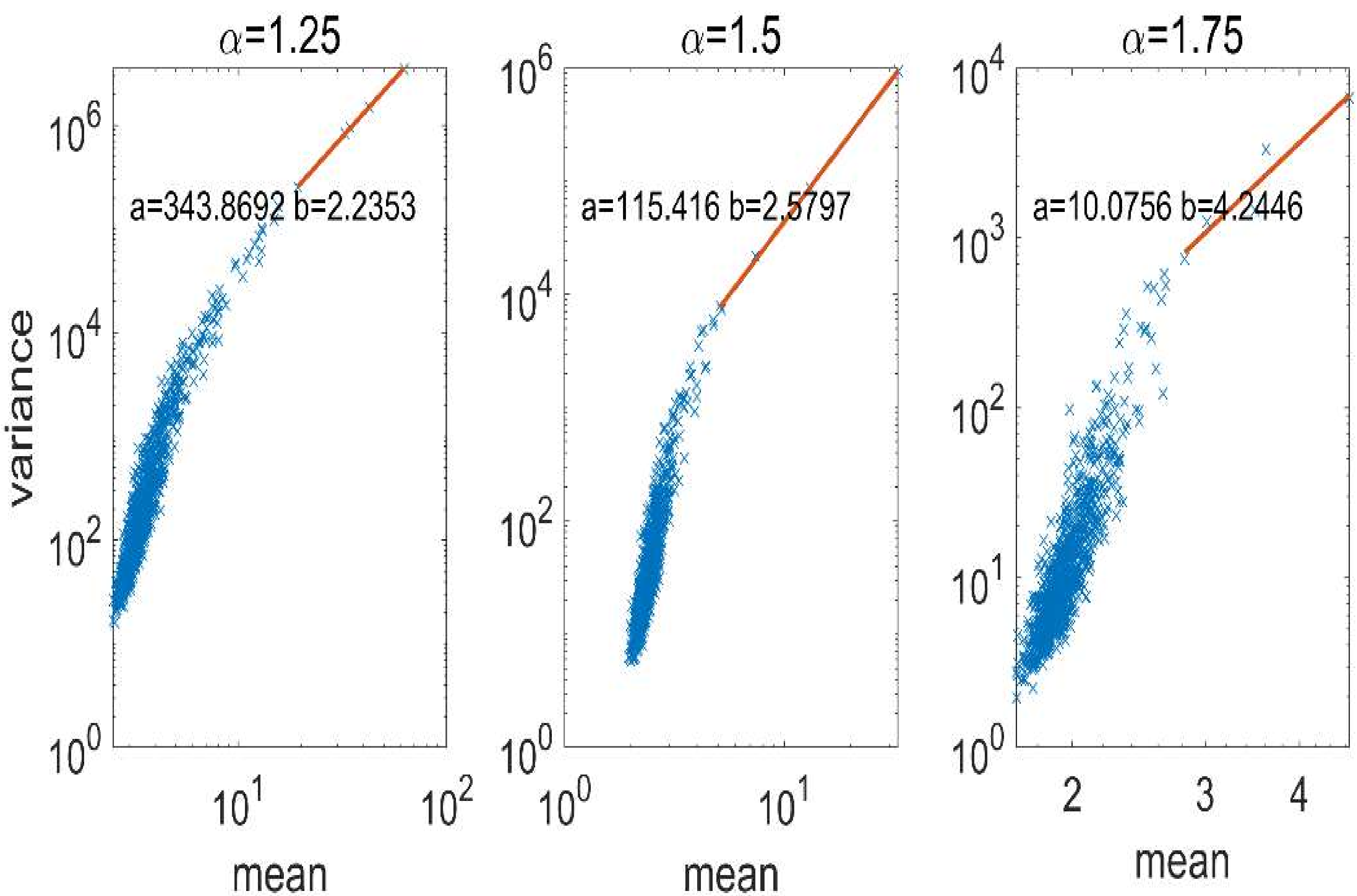
Lévy law: $\alpha = \frac{1}{2}, X = |N(0,1)|^{-2}$: straight line fitted to upper 0.25% of sampled means has slope ≈ 2.



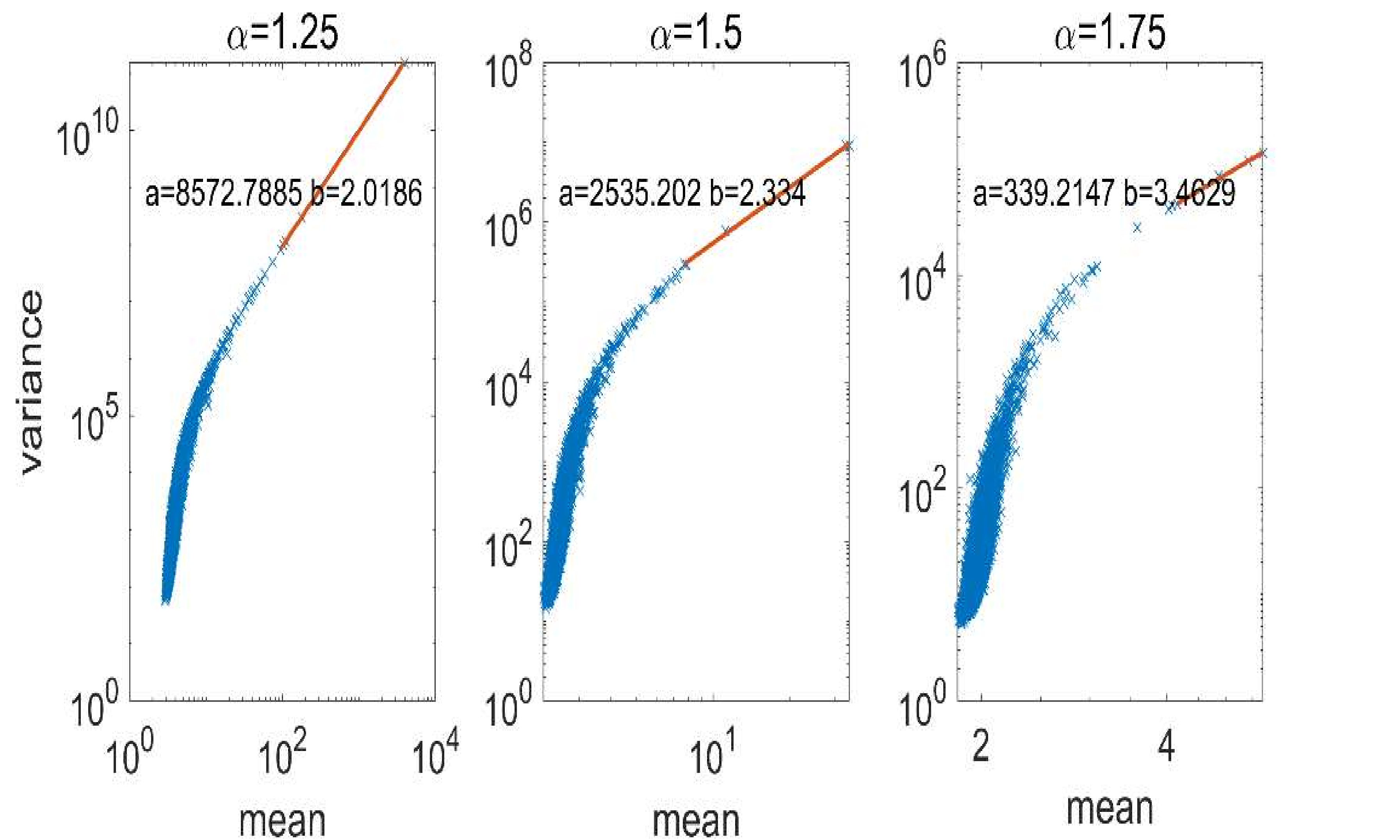
10^4 samples, each sample size 10^4, for $\alpha \in (0,1)$



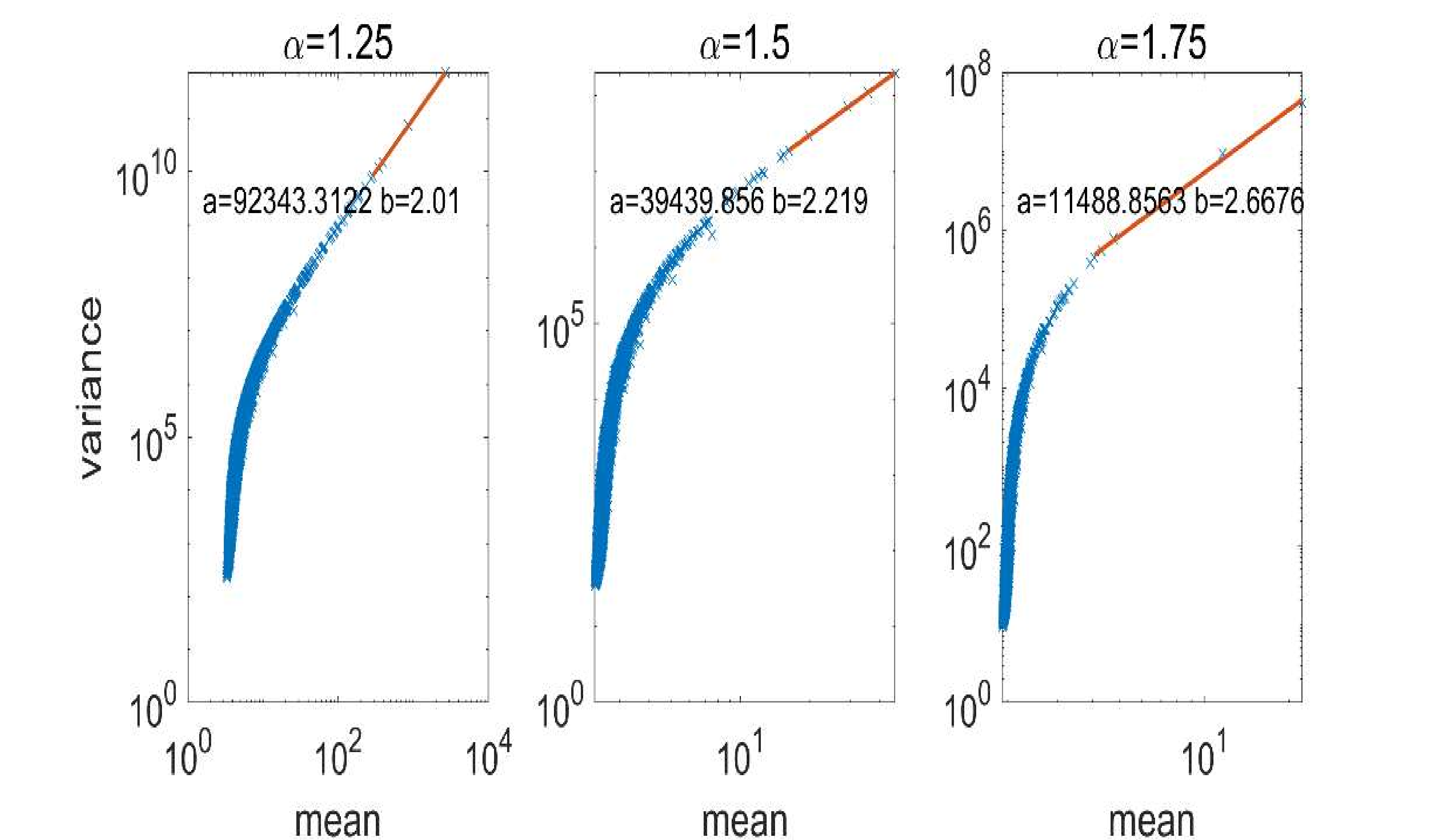
10^3 samples, each sample size 10^3, for $\alpha \in (1,2)$



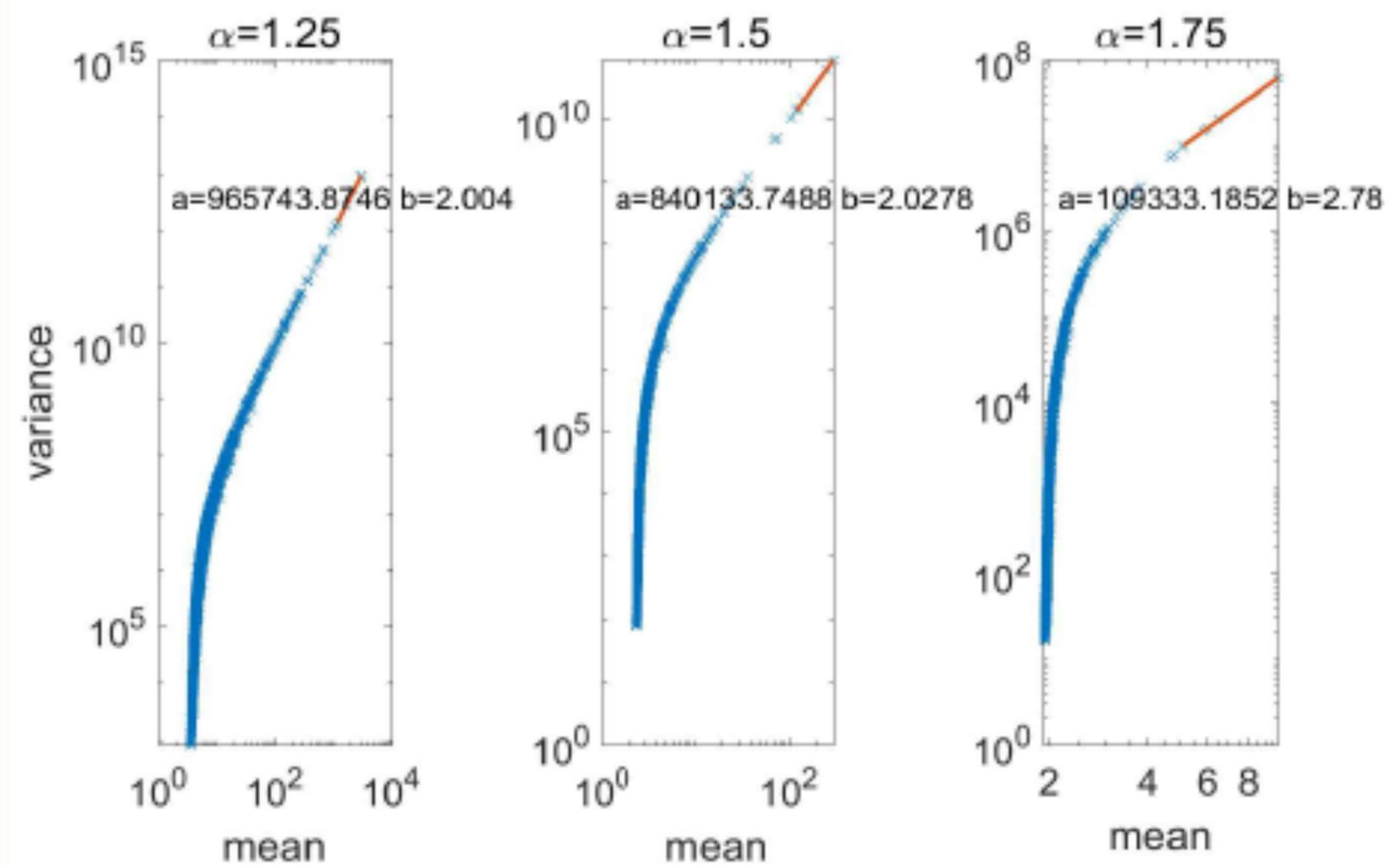
10^4 samples, each sample size 10^4, for $\alpha \in (1,2)$



10^5 samples, each sample size 10^5, for $\alpha \in (1,2)$



10^6 samples, each sample size 10^6, for $\alpha \in (1,2)$



Estimated slope *b* of log variance as a function of log mean

for 5 largest values of log mean in *c* samples, each of size *r* = *c*

	Tail index		
$r = c$	$\alpha = 1.25$	$\alpha = 1.5$	$\alpha = 1.75$
10^3	2.2353	2.5797	4.2446
10^4	2.0186	2.3340	3.4629
10^5	2.0100	2.2190	2.6676
10^6	2.0040	2.0278	2.7800

Outline

1. COVID-19 U.S. data analysis:
   Taylor's law & heavy upper tails
2. Simulations of an idealization
3. Mathematics

Upper tail of multiple samples obeys Taylor's law with slope 2.

Richard A. Davis

Gennady Samorodnitsky

TL holds when $0 < \alpha < 1$: w.h.p., $\log var(X_{:j}) \approx \log r + 2 \log \bar{X}_{:j}$

Theorem 4.1. *Let $\varepsilon > 0$. Let every element $X_{ij}, i = 1, \ldots, r; j = 1, \ldots, c$ of the array $\mathbf{X}$ have the same distribution in $RV(\alpha)$, $0 < \alpha < 1$, and assume that all columns are equally distributed. Assume also that the entries within each column are, conditionally on a σ-field $\mathcal{G}$, independent. Let $x(r)$ satisfy*

$$\lim_{r \to \infty} x(r) = \infty \qquad \text{and} \qquad \lim_{r \to \infty} r \Pr(X_{11} > x(r)) = 0. \tag{4.1}$$

Let the number c of columns depend on the number r of rows in such a way that the function $c = c(r)$ satisfies

$$\lim_{r \to \infty} c(r) r^2 \max_{i=1,\ldots,r} E\left(\left[\frac{1}{x(r)} \int_0^{x(r)} \Pr\left(X_{i,1} > x \mid \mathcal{G}\right) dx\right]^2\right) = 0. \tag{4.2}$$

Then

$$\lim_{r \to \infty} \Pr\left(\left|\log \frac{var(X_{:j})}{r(\bar{X}_{:j})^2}\right| > \varepsilon \text{ for some } j = 1, \ldots, c(r) \text{ such that } r|\bar{X}_{:j}| > x(r)\right) = 0. \tag{4.3}$$

Assumptions that all X_{ij} are iid, & that different columns are mutually independent, can be relaxed, & same TL holds.

Theorem 4.4. *Let $\{Z_{ij}\}$ be an iid space-time process with a regularly varying distribution with index $\alpha \in (0,1)$. Let $\{G_{ij}\}$ be a stationary space-time Gaussian process with mean 0 and covariance function $\gamma(\cdot,\cdot)$, i.e., $Cov(G_{ij}, G_{i'j'}) = \gamma(i-i', j-j')$. Given $\mathcal{G} = \sigma(G_{ij},\, i=1,\dots,r; j=1,\dots,c)$, the process $X_{ij} := Z_{ij}|G_{ij}|$ is conditionally independent, but not identically distributed. Under the assumptions (4.1) and (4.2), (4.3) holds.*

TL holds when $1 < \alpha < 2$: w.h.p., $\log var(X_{:j}) \approx \log r + 2 \log \bar{X}_{:j}$

Theorem 4.2. *Let $\varepsilon > 0$. Suppose the elements of the matrix are iid, with the right tail in $RV(\alpha)$, $1 < \alpha < 2$, and the left tail with a finite second moment. Let $x(r)$ satisfy*

$$\lim_{r\to\infty} x(r) = \infty \qquad \textit{and} \qquad \lim_{r\to\infty} r \Pr(X_{11} > x(r)) = 0. \tag{4.9}$$

Let the number c of columns depend on the number r of rows in such a way that the function $c = c(r)$ satisfies

$$\lim_{r\to\infty} c(r)\big(r \Pr(X_{11} > x(r))\big)^2 = 0. \tag{4.10}$$

Then

$$\lim_{r\to\infty} \Pr\left(\left|\log \frac{var(X_{:j})}{r(\bar{X}_{:j})^2}\right| > \varepsilon \textit{ for some } j = 1, \ldots, c(r) \textit{ such that } r\big(\bar{X}_{:j} - EX_{11}\big) > x(r)\right) = 0. \tag{4.11}$$

Relevant prior papers

Brown, M., Cohen, J. E. & de la Peña, V. 2017 Taylor's law, via ratios, for some distributions with infinite mean. Journal of Applied Probability 54(3):1-13. doi:10.1017/jpr.2017.25

Cohen, J. E., Davis, R.A. & Samorodnitsky, G. 2020 Heavy-tailed distributions, correlations, kurtosis, and Taylor's law of fluctuation scaling. Proceedings of the Royal Society A 476:20200610. https://doi.org/10.1098/rspa.2020.0610.

Thank you!
Questions?

cohen@rockefeller.edu

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

2021-09-15