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

\[ \text{variance} = f(\text{mean}) \]

Taylor's law: \( \text{variance} = a(\text{mean})^b \)
\[ \log(\text{variance}) = \log a + b \log(\text{mean}) \]

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

\[ \text{variance} = f(\text{mean}) \]

Taylor's law: \[ \text{variance} = a(\text{mean})^b \]
\[ \log(\text{variance}) = \log a + b \log(\text{mean}) \]

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: \( Var(X(s)) = f(E[X(s)]), \forall s \in S. \)

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

Fluctuation scaling, variance function

\[ \text{variance} = f(\text{mean}) \]

→ Taylor's law: \[ \text{variance} = a(\text{mean})^b \]
\[ \log(\text{variance}) = \log a + b \log(\text{mean}) \]

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

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

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

<table>
<thead>
<tr>
<th>Sample number</th>
<th>s=1</th>
<th>s=2</th>
<th>s=3</th>
<th>s=…</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size or density in units (quadrats, plots, transects, counties, states, years, days)</td>
<td>$x_{11}$</td>
<td>$x_{12}$</td>
<td>$x_{13}$</td>
<td>$x_{...}$</td>
</tr>
<tr>
<td></td>
<td>$x_{21}$</td>
<td>$x_{22}$</td>
<td>$x_{23}$</td>
<td>…</td>
</tr>
<tr>
<td></td>
<td>$x_{31}$</td>
<td>$x_{32}$</td>
<td>$x_{33}$</td>
<td>…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$x_{42}$</td>
<td>$x_{43}$</td>
<td>…</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$x_{52}$</td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>Mean (weighted)</td>
<td>$m_1$</td>
<td>$m_2$</td>
<td>$m_3$</td>
<td>$m_{...}$</td>
</tr>
<tr>
<td>Variance (weighted)</td>
<td>$v_1$</td>
<td>$v_2$</td>
<td>$v_3$</td>
<td>$v_{...}$</td>
</tr>
</tbody>
</table>
USA has more tornadoes than any other country. (Lloyd’s)
# TL for tornadoes: size of outbreaks by calendar year

<table>
<thead>
<tr>
<th>Year→</th>
<th>1954</th>
<th>1955</th>
<th>…</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of F1+ tornadoes per outbreak (≥6 consecutive tornadoes with ≤6 hour gap)</td>
<td>$x_{11}$</td>
<td>$x_{12}$</td>
<td>$x_{13}$</td>
<td>$x_{...}$</td>
</tr>
<tr>
<td></td>
<td>$x_{21}$</td>
<td>$x_{22}$</td>
<td>$x_{23}$</td>
<td>…</td>
</tr>
<tr>
<td></td>
<td>$x_{31}$</td>
<td>$x_{32}$</td>
<td>$x_{33}$</td>
<td>…</td>
</tr>
<tr>
<td>Mean (weighted)</td>
<td>$m_1$</td>
<td>$m_2$</td>
<td>$m_3$</td>
<td>$m_{...}$</td>
</tr>
<tr>
<td>Variance (weighted)</td>
<td>$v_1$</td>
<td>$v_2$</td>
<td>$v_3$</td>
<td>$v_{...}$</td>
</tr>
</tbody>
</table>
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

\[ b = 4.33 \pm 0.44 \]

Variance of # of tornadoes per outbreak \( \propto (\text{mean})^{4.33} \)
Higher percentiles increased faster. “quantile regression”

(a) Percentiles of tornadoes per outbreak

Tornadoes per outbreak over the years from 1960 to 2010, showing the percentiles of tornadoes per outbreak.

(b) Linear growth rate

Linear growth rate showing the trend from 1965 to 2015. The trend is given by $0.0055 \exp(4.27p)$.

Plan

Fluctuation scaling, variance function

\[ \text{variance} = f(\text{mean}) \]

Taylor's law: 
\[ \log(\text{variance}) = \log a + b \log(\text{mean}) \]

\[ \rightarrow \text{Heavy tails} \& \text{regular variation} \]

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

COVID-19 in US

Taylor's law, infinite variance
Lognormal distribution

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 \( \mu \), variance \( \sigma^2 \)).

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

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

If \( \sigma^2 \) is constant & only \( \mu \) changes, then

\[
\text{Var}(Y(\mu)) = c\left\{E(Y(\mu))\right\}^2 : \text{TL with exponent 2}.
\]
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$(0,1)$ pdf = $\frac{1}{\sqrt{2\pi x^2 e^{2x}}} \frac{1}{3} \cdot 1$
Lévy distribution has heavier right tail than lognormal distribution.

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

\[
\text{Lognormal}(0, 1) \text{ pdf } = \frac{1}{\sqrt{2\pi (\log x)^2/2}}
\]
Lévy cumulative averages grow like $n \times \text{Lévy}$. Lognormal averages converge.
Lévy cumulative averages grow like $n \times \text{Lévy}$. Lognormal averages converge.
Lévy law (stable $\alpha = 1/2$) obeys TL with increasing sample sizes.

Theoretical (not fitted) line of slope $b=3$

For stable law of index $\alpha \in (0,1)$, slope is $b = \frac{2-\alpha}{1-\alpha}$.

Brown, Cohen, de la Peña

J. Appl. Prob. 2017

Sample mean & sample variance diverge to $\infty$ as $n \to \infty$, but Taylor’s law holds!
Heterogeneous dependent data

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

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

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

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

$\log_{10} M_h$

$\log (\bar{X}_n)$, $n = 10^j$
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
\[ \text{variance} = f(\text{mean}) \]

Taylor's law: \[ \text{variance} = a(\text{mean})^b \]
\[ \log(\text{variance}) = \log a + b \log(\text{mean}) \]

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

<table>
<thead>
<tr>
<th>State number →</th>
<th>s=1</th>
<th>s=2</th>
<th>s=3</th>
<th>s=…</th>
</tr>
</thead>
<tbody>
<tr>
<td>County 1</td>
<td>$x_{11}$</td>
<td>$x_{12}$</td>
<td>$x_{13}$</td>
<td>$x_{...}$</td>
</tr>
<tr>
<td>County 2</td>
<td>$x_{21}$</td>
<td>$x_{22}$</td>
<td>$x_{23}$</td>
<td>$...$</td>
</tr>
<tr>
<td>County 3</td>
<td>$x_{31}$</td>
<td>$x_{32}$</td>
<td>$x_{33}$</td>
<td>$...$</td>
</tr>
<tr>
<td>County 4</td>
<td></td>
<td>$x_{42}$</td>
<td>$x_{43}$</td>
<td>$...$</td>
</tr>
<tr>
<td></td>
<td>$\vdots$</td>
<td></td>
<td>$x_{52}$</td>
<td>$...$</td>
</tr>
<tr>
<td>Mean = average</td>
<td>$m_1$</td>
<td>$m_2$</td>
<td>$m_3$</td>
<td>$m_{...}$</td>
</tr>
<tr>
<td>Variance</td>
<td>$v_1$</td>
<td>$v_2$</td>
<td>$v_3$</td>
<td>$v_{...}$</td>
</tr>
</tbody>
</table>
Variance function of cumulative U.S. COVID-19 cases/county by state

only states with ≥7 counties with >0 counts

Poisson

cases

slopes
(95% conf. interval)
only states with ≥7 counties with >0 counts
Taylor's law parameters of cumulative U.S. COVID-19 cases/county by state

$$s_n^2 = a \bar{X}_n^b$$

**Coefficient $a$**

**Exponent $b$**

---

Apr 2020 | Jul 2021
Apr 2020 | July 2021
---

**Cases**
Taylor's law parameters of cumulative U.S. COVID-19 deaths/county by state 

\[ s_n^2 = a \bar{X}_n^b \]

**Coefficient** \( a \)

**Exponent** \( b \)

- **Intercept** \( a \)
- **Slope** \( b \)

### Coefficient \( a \)
- April 2020: Approximately 2
- July 2021: Approximately 4

### Exponent \( b \)
- April 2020: Approximately 2.5
- July 2021: Approximately 2

**Month**
- April 2020
- July 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.
Survival curve of cumulative COVID-19 cases/county by date

Cases

$\Pr(X > x)$ log scale

$x$ log scale

Weibull

lognormal
Survival curve of cumulative COVID-19 deaths/county by date

Pr(X > x) log scale

lognormal

Weibull
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 $\sigma^2$ is constant, & if only $\mu$ varies from state to state, then Taylor's law holds with slope 2.
Lognormal $\sigma$ as function of $\mu$ for cumulative U.S. COVID-19 cases by state

state-specific lognormal $\sigma$

state-specific lognormal $\mu$
Deaths Lognormal $\sigma$ as function of $\mu$ for cumulative U.S. COVID-19 deaths by state
Lognormal explains TL with slope 2 for lower 99% of counts but not 1% upper tail.

Lognormal $\mu$ varies much more widely than lognormal $\sigma^2$, 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.
Survival curve of highest 1% of cumulative COVID-19 cases/county by date.
Survival curve of highest 1% of cumulative COVID-19 deaths/county by date

Pr(X > 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: $1 < \alpha < 2$
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$(\alpha)$: $|N(0,1)|^{-\alpha^{-1}}$, $|U|^{-\alpha^{-1}}$, $|U_1 U_2|^{-\alpha^{-1}}$, $|U_1 U_2 U_3|^{-\alpha^{-1}}$

$\alpha = 1/2, 1, 3/2$, with & without dependence. TL with $b \to 2$ is confirmed by mathematics.
4 models in RV(\(\alpha\)), 100 samples of size 100

\(\alpha = 1/2\)

\(\alpha = 1\)

\(\alpha = 1.5\)

Yes, we have theorems.

\[\log_{10} P(X > x)\]

\[\log_{10} x\]

\[\log_{10} \text{sample variance}\]

\[\log_{10} \text{sample mean}\]
$\rho = 0, 0.5, 0.9$, 100 samples of size 100

$\alpha = 0.5/2$

$\alpha = 1$

$\alpha = 1.5$

Yes, we have theorems.

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

log sample variance

log sample mean

log sample mean
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

Mark Brown  
Chuan-Fa Tang  
Richard A. Davis

Victor de la Peña  
Sheung Chi Yam  
Gennady Samorodnitsky
Thank you!
Questions?
cohen@rockefeller.edu

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

2/3/2024