Supporting Information

Cohen and Xu 10.1073/pnas.1503824112

SI Text

If X is a real-valued random variable with finite mean $E(X)$ and finite variance $var(X)$, and if a real-valued function f of real x is twice differentiable at $E(X)$, then the delta method (ref. 1 and ref. 2, pp. 355–358) gives the approximations

$$f(X) \approx f(E(X)) + (X - E(X))\left\{ \left(f'(x)\right)\big|_{x=E(X)} \right\},$$

$$E(f(X)) \approx f(E(X)) + \left\{ \frac{f''(x)}{2}\bigg|_{x=E(X)} \right\} \cdot var(X),$$

$$var(f(X)) \approx \left\{ \left(f'(x)\right)\big|_{x=E(X)} \right\}^2 var(X).$$

In practice, we compute sample moments from observations of X, plug them in to replace the population moments, and accept the result as approximations to the left sides.

Lemma 1. *Suppose Y is a nonnegative real-valued random variable with finite mean $E(Y) = M > 0$ and finite variance $var(Y) = V > 0$. Assume sampled observations are iid and the sample size in block j is n_j $(j = 1, 2, \ldots, N)$ and N is the number of blocks. If m_j is the sample mean of observations in block j, then the approximations given by the delta method are* $\log m_j \approx \log M + (m_j - M)/M$, $E(\log m_j) \approx \log M - V/(2n_j M^2)$, $var(\log m_j) \approx V/(n_j M^2)$.

Proof: In the approximations from the delta method, we set $X = m_j$, $f(x) = \log(x)$, $x > 0$. Therefore, $f'(x) = 1/x$ and $f''(x) = -1/x^2$. Because $E(m_j) = M$ and $var(m_j) = V/n_j$,

$$\log m_j \approx f(M) + (m_j - M) \cdot \frac{1}{M} = \log M + (m_j - M)/M,$$

$$E(\log m_j) \approx f(M) + \left(-\frac{1}{2M^2}\right) \cdot \frac{V}{n_j} = \log M - V/\left(2n_j M^2\right),$$

$$var(\log m_j) \approx \left(\frac{1}{M}\right)^2 \cdot \frac{V}{n_j} = V/\left(n_j M^2\right).$$

The proof is complete.

Lemma 2. *Under the assumptions of Lemma 1 also assume the third and fourth central moments of the random variable Y are finite and positive, that is, $\mu_h = E([Y - M]^h) > 0$, $h = 3, 4$. Suppose v_j is the unbiased sample variance of observations in block j and $E(v_j) = V > 0$. Then the approximations given by the delta method are* $\log v_j \approx \log V + (v_j - V)/V$, $var(\log v_j) \approx \left(\mu_4 - \frac{n_j - 3}{n_j - 1} V^2\right) \Big/ (n_j V^2)$, $E(\log v_j) \approx \log V - \frac{1}{2n_j}\left(\frac{\mu_4}{V^2} - \frac{n_j - 3}{n_j - 1}\right)$.

Proof: Setting $X = v_j$ and following the same arguments as in the proof of *Lemma 1* give the results.

Lemma 3. *Under the assumptions of Lemmas 1 and 2, the covariance of the sample mean and sample variance is* $cov(v_j, m_j) = \mu_3/n_j$.

Zhang (3) gives a proof of this classical formula, which has been known at least since 1903 (ref. 4, p. 279, equation xiii; ref. 5, p. 7, equation xxvi; ref. 6, p. 479, equation 67; and ref. 7, p. 402, equations 3 and 4).

Proof of Theorem: When all blocks are weighted equally, the least-squares estimators of slope b and intercept $\log(a)$, and SE of the slope estimator $s(\hat{b})$ are, respectively (8, p. 155),

$$\hat{b} = cov_+(\log v_j,\ \log m_j)/var_+(\log m_j),$$

$$\widehat{\log(a)} = mean_+(\log v_j) - \hat{b} \cdot mean_+(\log m_j),$$

$$s\left(\hat{b}\right) = \sqrt{\left[var_+(\log v_j)/var_+(\log m_j) - \left\{cov_+(\log v_j, \log m_j)\right\}^2 / \left\{var_+(\log m_j)\right\}^2\right]/(N-2)}.$$

The notations $mean_+(\cdot)$, $var_+(\cdot)$, and $cov_+(\cdot, \cdot)$ are to be read as the mean, variance, and covariance across all blocks and not as referring to any single block j. Explicitly, the sample estimators are defined by

$$mean_+(\log m_j) = \frac{1}{N}\sum_{j=1}^{N} \log m_j,$$

$$mean_+(\log v_j) = \frac{1}{N}\sum_{j=1}^{N}\log v_j,$$

$$var_+(\log m_j) = \frac{1}{N-1}\sum_{j=1}^{N}(\log m_j)^2 - \frac{1}{N(N-1)}\left(\sum_{j=1}^{N}\log m_j\right)^2,$$

$$var_+(\log v_j) = \frac{1}{N-1}\sum_{j=1}^{N}(\log v_j)^2 - \frac{1}{N(N-1)}\left(\sum_{j=1}^{N}\log v_j\right)^2,$$

$$cov_+(\log v_j, \log m_j) = \frac{1}{N-1}\sum_{j=1}^{N}(\log m_j \cdot \log v_j) - \frac{1}{N(N-1)}\left(\sum_{j=1}^{N}\log m_j\right)\left(\sum_{j=1}^{N}\log v_j\right).$$

They are all consistent by the law of large numbers: as $N \to \infty$, $mean_+(\log m_j) \to_P E(\log m_j)$, $mean_+(\log v_j) \to_P E(\log v_j)$, $var_+(\log m_j) \to_P var(\log m_j)$, $var_+(\log v_j) \to_P var(\log v_j)$, and $cov_+(\log v_j, \log m_j) \to_P cov(\log v_j, \log m_j)$. Here the symbol "$\to_P$" means convergence in probability.

To find the limits in probability of $\hat{b}$ and $s(\hat{b})$, we approximate the above estimators by the delta method using *Lemmas 1*, *2*, and *3*. We first approximate the numerator and the denominator of $\hat{b}$ separately. For the numerator of $\hat{b}$, namely, $cov_+(\log v_j, \log m_j)$, the first term is approximately

$$\begin{aligned}\frac{1}{N-1}\sum_{j=1}^{N}(\log m_j \cdot \log v_j) &\approx \frac{1}{N-1}\sum_{j=1}^{N}\left\{\log M + \frac{1}{M}(m_j - M)\right\}\cdot\left\{\log V + \frac{1}{V}(v_j - V)\right\}\\ &= \frac{N}{N-1}\cdot\log M\cdot\log V + \frac{\log V}{(N-1)M}\sum_{j=1}^{N}(m_j - M) + \frac{\log M}{(N-1)V}\sum_{j=1}^{N}(v_j - V) + \frac{1}{(N-1)MV}\sum_{j=1}^{N}(m_j - M)(v_j - V).\end{aligned}$$

The second term of the numerator of $\hat{b}$ is approximately

$$\begin{aligned}\frac{1}{N(N-1)}\left(\sum_{j=1}^{N}\log m_j\right)\left(\sum_{j=1}^{N}\log v_j\right) &\approx \frac{1}{N(N-1)}\sum_{j=1}^{N}\left\{\log M + \frac{1}{M}(m_j - M)\right\}\cdot\sum_{j=1}^{N}\left\{\log V + \frac{1}{V}(v_j - V)\right\}\\ &= \frac{N}{N-1}\cdot\log M\cdot\log V + \frac{\log V}{(N-1)M}\sum_{j=1}^{N}(m_j - M) + \frac{\log M}{(N-1)V}\sum_{j=1}^{N}(v_j - V)\\ &\quad + \frac{1}{N(N-1)MV}\sum_{j=1}^{N}(m_j - M)\sum_{j=1}^{N}(v_j - V).\end{aligned}$$

Therefore

$$\begin{aligned}cov_+(\log v_j, \log m_j) &\approx \frac{1}{(N-1)MV}\sum_{j=1}^{N}(m_j - M)(v_j - V) - \frac{1}{N(N-1)MV}\sum_{j=1}^{N}(m_j - M)\sum_{j=1}^{N}(v_j - V)\\ &= \frac{1}{(N-1)MV}\sum_{j=1}^{N}m_j v_j - \frac{1}{N(N-1)MV}\sum_{j=1}^{N}m_j\sum_{j=1}^{N}v_j = \frac{cov_+(m_j, v_j)}{MV}.\end{aligned}$$

Similarly, the denominator of $\hat{b}$ is approximately

$$var_+(\log m_j) \approx \frac{1}{M^2}\left\{\frac{1}{(N-1)}\sum_{j=1}^{N}m_j^2 - \frac{1}{N(N-1)}\left(\sum_{j=1}^{N}m_j\right)^2\right\} = var_+(m_j)/M^2.$$

Consequently, for large n_j, $j = 1, 2, \ldots, N$, $\hat{b} \approx \frac{cov_+(m_j, v_j)}{MV}\Big/\frac{var_+(m_j)}{M^2}$. By consistency, for large N, using *Lemma 3* in the numerator,

$$\hat{b} \approx \frac{cov(m_j, v_j)}{MV}\Bigg/\frac{var(m_j)}{M^2} = \frac{\mu_3}{n_j MV}\Bigg/\frac{V}{n_j M^2} = \mu_3 M/V^2 = \gamma_1/CV.$$

Using the consistency of estimator $mean_+(\cdot)$ and existing expressions for $E(\log m_j)$, $E(\log v_j)$ and $\hat{b}$, for large N and n_j, $j = 1, 2, \ \ldots, \ N$,

$$\widehat{\log(a)} \approx E(\log v_j) - \hat{b}\cdot E(\log m_j) \approx \left[\log V - \frac{1}{2n_j}\left(\frac{\mu_4}{V^2} - \frac{n_j-3}{n_j-1}\right)\right] - \frac{\gamma_1}{CV}\left[\log M - V/\left(2n_jM^2\right)\right] \approx \log V - \frac{\gamma_1}{CV}\cdot \log M.$$

The derivation of $var_+(\log v_j)$ is the same as that of $var_+(\log m_j)$. Replacing m_j with v_j and M with V yields $var_+(\log v_j) \approx var_+(v_j)/V^2$. For large N and $n_j, j=1, 2, \ldots, N$, substituting into the formula for $s(\hat{b})$ the estimators corresponding to $var_+(m_j)$, $var_+(v_j)$, and $\hat{b}$ yields

$$s\left(\hat{b}\right) \approx \sqrt{\frac{1}{N-2}\left[\left(\frac{\mu_4}{V^2}-1\right)\Big/\frac{V}{M^2} - \left(\mu_3 M/V^2\right)^2\right]} = \sqrt{\frac{M^2\left(\mu_4 V - V^3 - \mu_3^2\right)}{(N-2)V^4}} = \sqrt{\frac{\kappa - 1 - \gamma_1^2}{(N-2)(CV)^2}},$$

where $\kappa = \mu_4/V^2$ is the kurtosis. This completes the proof.

1. Oehlert GW (1992) A note on the delta method. *Am Stat* 46(1):27–29.
2. Hosmer DW, Lemeshow S, May S (2008) *Applied Survival Analysis: Regression Modeling of Time-to-Event Data* (Wiley, New York), 2nd Ed.
3. Zhang L (2007) Sample mean and sample variance: their covariance and their (in)dependence. *Am Stat* 61(2):159–160.
4. Pearson K (1903) On the probable errors of frequency constants. *Biometrika* 2(3):273–281.
5. Pearson K (1913) On the probable errors of frequency constants part II. *Biometrika* 9(1/2):1–10.
6. Neyman J (1925) Contributions to the theory of small samples drawn from a finite population. *Biometrika* 17(3/4):472–479.
7. Neyman J (1926) On the correlation of the mean and the variance in samples drawn from an "infinite" population. *Biometrika* 18(3/4):401–413.
8. Snedecor GW, Cochran WG (1980) *Statistical Methods* (Iowa State Univ Press, Ames, IA), 7th Ed.

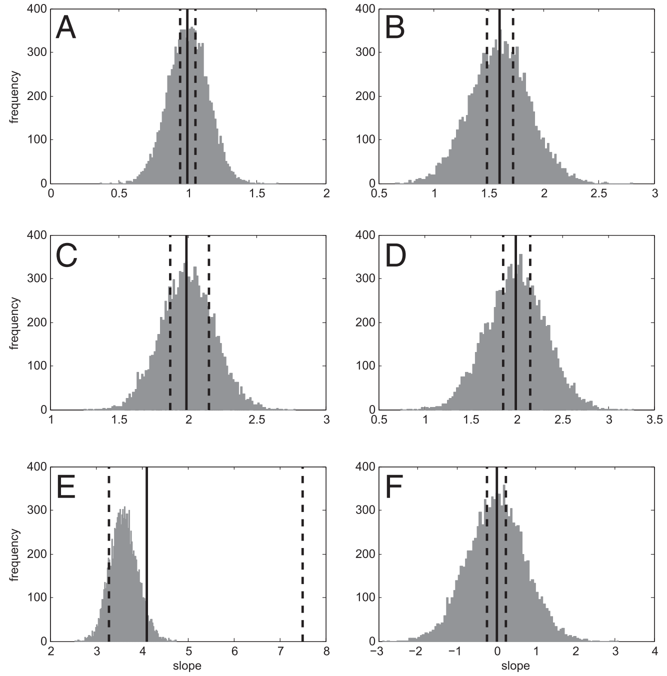

Fig. S1. Comparison of TL slope estimator $\hat{b}$ predicted from theory and computed using linear regression for (*A*) Poisson ($\lambda = 1$), (*B*) negative binomial ($r = 5$, $p = 0.4$), (*C*) exponential ($\lambda = 1$), (*D*) gamma ($\alpha = 4$, $\beta = 1$), (*E*) lognormal ($\mu = 1$, $\sigma = 1$), and (*F*) shifted normal [$5 + \mathcal{N}(0,1)$] distributions. Gray histogram shows the distribution of point estimates of b from 10,000 linear regressions. For each distribution, the black solid line and dashed lines give, respectively, the median and 95% CI of b calculated from 10,000 random copies of $n \times N$ iid samples using the theoretical formula (Eq. **3**).

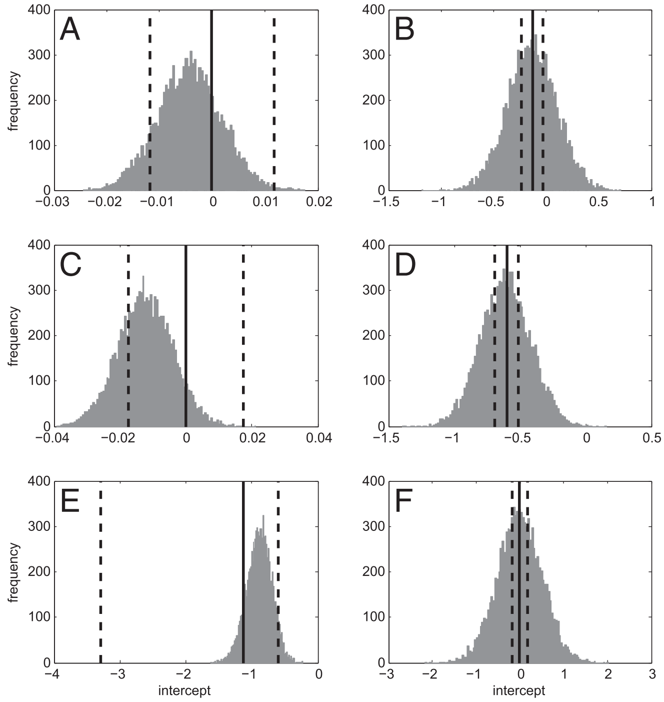

Fig. S2. Comparison of TL intercept estimator $\widehat{\log(a)}$ predicted from theory and computed using linear regression for (*A*) Poisson ($\lambda = 1$), (*B*) negative binomial ($r = 5$, $p = 0.4$), (*C*) exponential ($\lambda = 1$), (*D*) gamma ($\alpha = 4$, $\beta = 1$), (*E*) lognormal ($\mu = 1$, $\sigma = 1$), and (*F*) shifted normal $[5 + \mathcal{N}(0,1)]$ distributions. Gray histogram shows the distribution of point estimates of log(*a*) from 10,000 linear regressions. For each distribution, the black solid line and dashed lines gave, respectively, the median and 95% CI of log(*a*) calculated from 10,000 random copies of $n \times N$ iid samples using the theoretical formula (Eq. **4**).

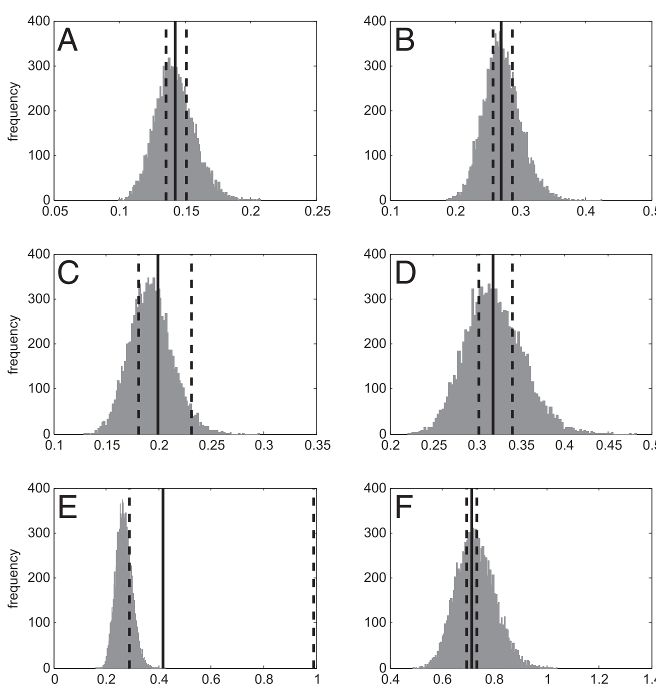

Fig. S3. Comparison of SE of the slope estimator $[s(\hat{b})]$ predicted from theory and computed using linear regression for (*A*) Poisson ($\lambda = 1$), (*B*) negative binomial ($r = 5$, $p = 0.4$), (*C*) exponential ($\lambda = 1$), (*D*) gamma ($\alpha = 4$, $\beta = 1$), (*E*) lognormal ($\mu = 1$, $\sigma = 1$), and (*F*) shifted normal $[5 + \mathcal{N}(0,1)]$ distributions. Gray histogram shows the distribution of point estimates of the SE of $\hat{b}$ from 10,000 linear regressions. For each distribution, the black solid line and dashed lines gave, respectively, the median and 95% CI of the SE of $\hat{b}$ calculated from 10,000 random copies of $n \times N$ iid samples using the theoretical formula (Eq. **5**).