Supplementary Material: Finite Sample Corrections for Parameters Estimation and Significance Testing

1 MATHEMATICAL DERIVATIONS RELATED TO POWER-LAW AND EXPONENTIAL DISTRIBUTIONS

The following are the notations use in this Section:

Table S1. Notations use in Section: Mathematical derivations related to Power Law (PL) and Exponential Distribution (EXP)

Notation	Description	Example: PL, EXP
$f_*(x)$	Probability density function (PDF) for $*$ distribution	$f_{PL}(x)$, $f_{EXP}(x)$
$F_*(x)$	Cumulative density function (CDF) for $*$ distribution	$F_{PL}(x)$, $F_{EXP}(x)$
$F_*^{-1}(u)$	Inverse cumulative density function for $*$ distribution	$F_{PL}^{-1}(u)$, $F_{EXP}^{-1}(u)$
$\mathbb{L}_*$	Likelihood function for $*$ distribution	$\mathbb{L}_{PL}$, $\mathbb{L}_{EXP}$
$\hat{P}*$	Estimated parameter(s) for $*$ distribution	$\hat{\alpha}$, $\hat{\beta}$
P_T^*	True parameter(s) for $*$ distribution	α_T, β_T
RE_*	Relative estimation error, $\frac{\sqrt{\langle(\hat{P}*-P_T^*)^2\rangle}}{P_T^*}$	RE_{PL}, RE_{EXP}
x_{max}	Postulated largest element, $x_{max} = F^{-1}(1-\delta)$	X_{max}^{PL}, X_{max}^{EXP}

Review of Probability Distributions and Parameter Estimation

The probability density functions (PDFs) for power-law (PL) and exponential (EXP) distributions are

$$f_{PL}(x) = \frac{\alpha-1}{x_{min}^{1-\alpha}}x^{-\alpha}; \qquad f_{EXP}(x) = \beta\exp(-\beta(x-x_{min})) \tag{S1}$$

for $x \in [x_{min}, \infty)$, and their cumulative distribution functions (CDFs) are

$$F_{PL}(x) = 1 - \left(\frac{x}{x_{min}}\right)^{1-\alpha}; \qquad F_{EXP}(x) = 1-\exp(-\beta(x-x_{min})). \tag{S2}$$

Hence, their inverse CDFs are

$$F_{PL}^{-1} = x_{min}(1-u)^{\frac{1}{1-\alpha}}; \qquad F_{EXP}^{-1} = x_{min} - \frac{\ln(1-u)}{\beta}, \tag{S3}$$

which maps $u \sim U(0,1)$ to $x \in [x_{min}, \infty)$ for the PL and EXP distributions respectively.

We use the maximum likelihood method to estimate the parameters in the distribution. To do so, we write down the log-likelihood function:

$$\ln\mathbb{L} = \ln\left[\prod_{i=1}^{N} f\left(x_i|\hat{P}\right)\right]:$$

$$\ln\mathbb{L}_{PL} = N\left(\ln(\alpha-1) - \ln(x_{min})\right) - \alpha\sum_{i=1}^{N}\ln\left(\frac{x_i}{x_{min}}\right), \qquad \ln\mathbb{L}_{EXP} = N\ln(\beta) - \beta\sum_{i=1}^{N}(x_i - x_{min}) \tag{S4}$$

for both distributions. Applying the maximization condition $\frac{\partial[\ln L])}{\partial\alpha} = 0$ for the PL and $\frac{\partial[\ln L]}{\partial\beta} = 0$ for the EXP distributions, yield the estimated parameters,

$$\hat{\alpha} = 1 + \frac{1}{\langle \ln\left(\frac{x}{x_{min}}\right)\rangle}; \qquad \hat{\beta} = \frac{1}{\langle x\rangle - x_{min}}. \tag{S5}$$

where the angled brackets denote the mean. Note that the above condition is necessary but not sufficient to show that $\mathbb{L}$ is maximized. In addition, we also can see that the sufficient condition $\frac{\partial^2[\ln \mathbb{L}]}{\partial\alpha^2} = -\frac{1}{(\hat{\alpha}-1)^2} < 0$ and $\frac{\partial^2[\ln \mathbb{L}]}{\partial\beta^2} = -\frac{1}{\hat{\beta}^2} < 0$ are also satisfied by all $\hat{\alpha}$ and $\hat{\beta}$.

Adjustment for Finite Largest Element

In the main paper, we have demonstrated how the finite largest element effect (FLE) causes us to underestimate the value of $\langle \ln(x)\rangle$ (for PL) and $\langle x\rangle$ (for EXP). This results in an overestimated $\hat{\alpha}$ (for PL) and $\hat{\beta}$ (for EXP). We correct for this by adding back the truncated parts (from x_{max} to ∞) into $\langle \ln(x)\rangle$ and $\langle x\rangle$,

$$\begin{gathered}\left\langle \ln\left(\frac{x}{x_{min}}\right)\right\rangle_{adj} = \left\langle \ln\left(\frac{x}{x_{min}}\right)\right\rangle_{data} \int_{x_{min}}^{x_{max}} f_{PL}(x)\,dx + \int_{x_{max}}^{\infty} \ln\left(\frac{x}{x_{min}}\right) f_{PL}(x)\,dx; \\ \langle x\rangle_{adj} = \langle x\rangle_{data} \int_{x_{min}}^{x_{max}} f_{EXP}(x)\,dx + \int_{x_{max}}^{\infty} x f_{EXP}(x)\,dx.\end{gathered} \tag{S6}$$

This yields the adjusted values,

$$\begin{gathered}\left\langle \ln\left(\frac{x}{x_{min}}\right)\right\rangle_{adj} = \left\langle \ln\left(\frac{x}{x_{min}}\right)\right\rangle_{data} \left[1 - \left(\frac{x_{max}}{x_{min}}\right)^{1-\alpha}\right] + \frac{1}{\alpha-1}\left(\frac{x_{max}}{x_{min}}\right)^{1-\alpha}\left[(\alpha-1)\ln\left(\frac{x_{max}}{x_{min}}\right)+1\right] \\ \langle x\rangle_{adj} = \langle x\rangle_{data}\left[1 - \exp\left(-\beta(x_{max}-x_{min})\right)\right] + \frac{1}{\beta}\left(\beta x_{max}+1\right)\exp\left(-\beta(x_{max}-x_{min})\right),\end{gathered} \tag{S7}$$

for the PL and EXP distributions respectively. By inserting Equation (S7) into Equation (S5), we obtain the nonlinear equations

$$\begin{gathered}\left\langle \ln\left(\frac{x}{x_{min}}\right)\right\rangle_{data} + \frac{1}{1-\hat{\alpha}_{adj}} + \left(\frac{x_{max}}{x_{min}}\right)^{1-\hat{\alpha}_{adj}}\left[\ln\left(\frac{x_{max}}{x_{min}}\right) - \left\langle \ln\left(\frac{x}{x_{min}}\right)\right\rangle_{data} + \frac{1}{\hat{\alpha}_{adj}-1}\right] = 0; \\ \left[\hat{\beta}_{adj}\left(x_{max} - \langle x\rangle_{data}\right) + 1\right]\exp\left[-\hat{\beta}_{adj}\left(x_{max}-x_{min}\right)\right] + \hat{\beta}_{adj}\left(\langle x\rangle_{data} - x_{min}\right) - 1 = 0\end{gathered} \tag{S8}$$

that can be solved to obtain the adjusted estimated $\hat{\alpha}$ and $\hat{\beta}$.

Approximate the Relative Error

To estimate the errors in P^* arising due to the FLE effect, x_{max}, we define the relative estimated parameter error as

$$RE = \frac{\sqrt{\langle\left(\hat{P} - P_T\right)^2\rangle}}{P_T}. \tag{S9}$$

The finite largest element x_{max} can be expressed in term of δ using Equation (S3),

$$X_{max}^{PL} = x_{min}\delta^{\frac{1}{\alpha_T - 1}}; \qquad X_{max}^{EXP} = x_{min} - \frac{\ln(\delta)}{\beta_T}, \tag{S10}$$

the unadjusted terms $\langle \ln(x) - \ln(x_{min}) \rangle_{unadj}$ and $\langle x \rangle_{unadj}$ only contain $x \in [x_{min}, x_{max}]$, so they can be expressed as,

$$\left\langle \ln\left(\frac{x}{x_{min}}\right)\right\rangle_{unadj} = \int_{x_{min}}^{X_{max}^{PL}} \ln\left(\frac{x}{x_{min}}\right) f_{PL}\left(x|\alpha_T\right) dx; \qquad \langle x \rangle_{unadj} = \int_{x_{min}}^{X_{max}^{EXP}} x f_{EXP}\left(x|\beta_T\right) dx. \tag{S11}$$

Evaluating the integral for both equations gives,

$$\begin{aligned} \left\langle \ln\left(\frac{x}{x_{min}}\right)\right\rangle_{unadj} &= \frac{1}{\alpha_T - 1}\left\{1 - \left(\frac{X_{max}^{PL}}{x_{min}}\right)^{1-\alpha_T}\left[(\alpha_T - 1)\ln\left(\frac{X_{max}^{PL}}{x_{min}}\right) + 1\right]\right\} \\ \langle x \rangle_{unadj} &= \frac{1}{\beta_T}\left\{(\beta_T x_{min} + 1) - \left(\beta_T X_{max}^{EXP} + 1\right)\exp\left(-\beta_T(\,X_{max}^{EXP} - x_{min})\right)\right\}. \end{aligned} \tag{S12}$$

To check the impact this has on the estimated parameters, we substitute Equation (S12) into Equation (S5) to obtain the unadjusted estimated parameter for PL distribution,

$$\begin{aligned} \hat{\alpha}_{unadj} &= 1 + \frac{\alpha_T - 1}{1 - \left(\frac{X_{max}^{PL}}{x_{min}}\right)^{1-\alpha_T}\left[(\alpha_T - 1)\ln\left(\frac{X_{max}^{PL}}{x_{min}}\right) + 1\right]} \\ &\approx \alpha_T + (\alpha_T - 1)\left(\frac{X_{max}^{PL}}{x_{min}}\right)^{1-\alpha_T}\left[(\alpha_T - 1)\ln\left(\frac{X_{max}^{PL}}{x_{min}}\right) + 1\right] \\ &\quad + \mathcal{O}\left\{\left(\frac{X_{max}^{PL}}{x_{min}}\right)^{2(1-\alpha_T)}\left[(\alpha_T - 1)\ln\left(\frac{X_{max}^{PL}}{x_{min}}\right) + 1\right]^2\right\} \end{aligned} \tag{S13}$$

and for EXP distribution,

$$\begin{aligned} \hat{\beta}_{unadj} &= \frac{\beta_T}{1 - (\beta_T X_{max}^{EXP} + 1)\exp\left(-\beta_T(\,X_{max}^{EXP} - x_{min})\right)} \\ &\approx \beta_T + \beta_T\left[\beta_T X_{max}^{EXP} + 1\right]\exp\left(-\beta_T(X_{max}^{EXP} - x_{min})\right) \\ &\quad + \mathcal{O}\left\{\left[\beta_T X_{max}^{EXP} + 1\right]^2\exp\left(-2\beta_T(X_{max}^{EXP} - x_{min})\right)\right\}. \end{aligned} \tag{S14}$$

The approximations in Equation (S13) and Equation (S14) are estimated using first order Taylor expansion. Rewriting $\hat{\alpha}_{unadj}$ and $\hat{\beta}_{unadj}$ in terms of δ from Equation (S10) will lead to

$$\begin{aligned} \alpha_{unadj} &\approx \alpha_T + (\alpha_T - 1)\,\delta\left[1 - \ln(\delta)\right] \\ \beta_{unadj} &\approx \beta_T + \beta_T\delta\left[1 - \ln(\delta) + \beta_T x_{min}\right]. \end{aligned} \tag{S15}$$

Rewriting the relative error for PL and EXP distribution in term of δ, we get

$$\begin{aligned} RE_{PL}(\alpha_{unadj}) &= \left(1 - \frac{1}{\alpha_T}\right)\delta\left[1 - \ln(\delta)\right], \qquad \delta \in [0, 1] \\ RE_{EXP}(\beta_{unadj}) &= \delta\left[1 - \ln(\delta) + \beta_T x_{min}\right], \qquad \delta \in [0, 1]. \end{aligned} \tag{S16}$$

While it is useful to note that both relative errors are monotonic increasing functions with $RE_{PL} \in [0, 1]$ and $RE_{EXP} \in [0, \beta_T x_{min}]$, it is more important to observe that for a given δ, RE_{PL} is always smaller or equal to RE_{EXP}. This sensitivity of the estimated parameter (relative error) on finite largest element ($x_{max}(\delta)$) provides us an insight on answering why it is easier for the PL distribution to pass significance testing compared to the EXP distribution using unadjusted parameters.

2 RELATIONSHIP BETWEEN KS DISTANCE, DISTRIBUTION NOISE, AND SAMPLE SIZE

For a reordered data set $X = \{x_1^{(em)}, x_2^{(em)}, \ldots, x_N^{(em)}\}$ such that $x_i^{(em)} < x_j^{(em)}$ for $i < j$. If X is f_x distributed and has CDF F_x, the probability integral transform transforms X to the standard uniform distribution. Mathematically, this is the transformation $X = \{x_1^{(em)}, x_2^{(em)}, \ldots x_N^{(em)}\} \rightarrow \{F_x^{-1}(x_1^{(em)}) = u_1^{(em)}, F_x^{-1}(x_2^{(em)}) = u_2^{(em)}, \ldots, F_x^{-1}(x_N^{(em)}) = u_N^{(em)}\}$.[1] Thereafter, statistical significance testing is done by examining how plausible the empirically transformed set $U^{(em)} = \{u_1^{(em)}, u_2^{(em)}, \ldots u_N^{(em)}\} = U^{(em)}$ can be obtained by chance, from the standard uniform distribution, $U^{(s)} = \{u_1^{(s)}, u_2^{(s)}, \ldots u_N^{(s)}\}$. This can be done by comparing the KS distance and distribution noise value between $U^{(em)}$ and $U^{(s)}$. Since $U^{(em)}$ is fixed and depends only on the sample, we only need to focus on the samples drawn from standard uniform distribution ($U^{(s)}$) to understand the relationship between KS distance (d_{KS}), distribution noise (d_{DN}) and sample size (N).

We simulated 10^6 samples of various sizes N from $U(0,1)$. Thus, for each N we have 10^6 simulated sample $U^{(s)} = \{u_1^{(s)}, u_1^{(s)}, \ldots, u_N^{(s)}\}$ for $s = [1, 10^6]$ ordered such that $u_i^{(s)} \leq u_{i+1}^{(s)}$. For each sample then we calculate its KS distance d_{KS} as

$$d_{KS} = \forall_{i=1}^{N} \sup \left(\left| u_i - \frac{i}{N} \right| \right), \tag{S17}$$

and also its distribution noise d_{DN} using

$$d_{DN} = \sqrt{\frac{\sum_{i=1}^{N} (u_i - u_{i-1})^2 \left[f(u_{i-1}, u_i) - 1\right]^2}{\sum_{i=1}^{N} (u_i - u_{i-1})^2}} = \sqrt{\frac{\sum_{i=1}^{N} (u_i - u_{i-1})^2 \left(\frac{1}{N(u_i - u_{i-1})} - 1\right)^2}{\sum_{i=1}^{N} (u_i - u_{i-1})^2}},$$

Hence, for each sample size N, we will obtain 10^6 pairs of d_{KS} and d_{DN}.

2.1 KS Distance and Sample Size

The d_{KS} value at different deciles exhibits the power law relationship as $d_{KS} \propto A_{KS} N^{-\alpha_{KS}}$ for $N > 50$, shown in Figure S1. By assuming that the proportionality constant A_{KS} depends only on the percentile level $\wp_{KS}$, we begin experimenting with several functional forms and eventually settled for

$$A_{KS} = \left(\frac{100}{\wp_{KS}} - 1 \right)^{-0.176} \exp(-0.274). \tag{S18}$$

The value of the exponent α_{KS} can be determined by the slopes of the different deciles. For simplicity, we assume constant exponent,

$$\alpha_{KS} = 0.492, \tag{S19}$$

which is the mean of the slopes. In Figure S2, we fit the A_{KS} and α_{KS} to the different percentiles $\wp_{KS}$, which provided insight when choosing the functional form for α_{KS} and A_{KS}. Combining Equation (S18) with Equation (S19), we obtain the d_{KS} and N at percentile $\wp_{KS}$,

$$d_{KS}(\wp_{KS}, N) = \frac{\left(\frac{100}{\wp_{KS}} - 1 \right)^{-0.176} \exp(-0.274)}{N^{0.492}}, \quad N > 50. \tag{S20}$$

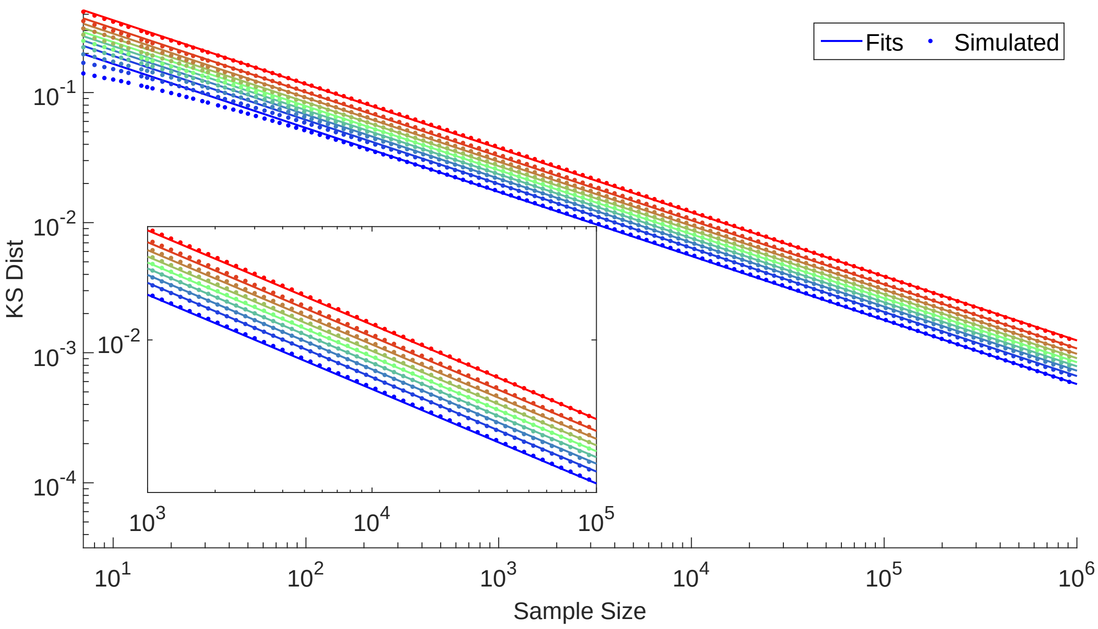

Figure S1: Log-log plot of d_{KS} against N for different deciles going from the 10^{th} percentile (blue) to the 90^{th} (red), obtained from 10^6 simulations.

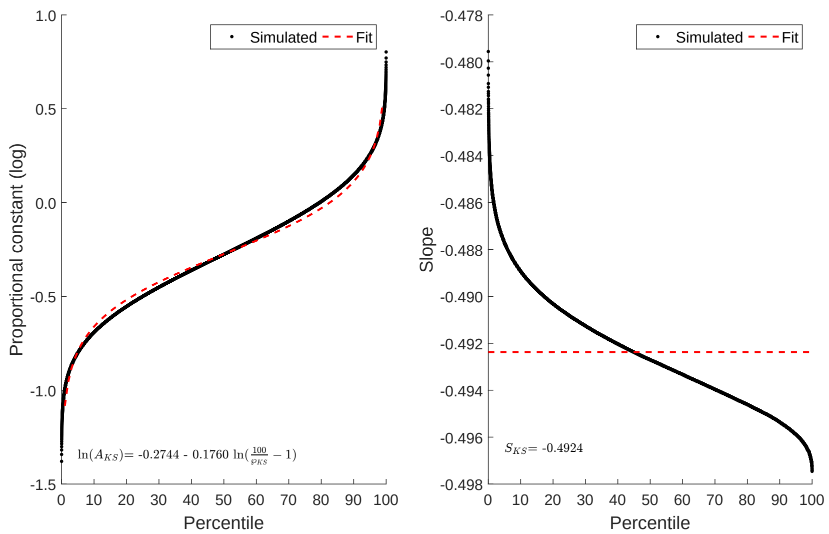

Figure S2: Fits for A_{KS} and α_{KS} at different percentile.

2.1.1 Analytical Solution Relating Distribution Noise to Sample Size

For a sample distributed by $U(0,1)$ we are able to sort it to get $U^{(s)}$, where the superscript (s) refers to the ordered sample. We now define $\Delta = \{\Delta_1 = u_1^{(s)}, \Delta_2 = u_2^{(s)} - u_1^{(s)}, \Delta_3 = u_3^{(s)} - u_2^{(s)}, \ldots, \Delta_N =$

$u_N^{(s)} - u_{N-1}^{(s)}\}$. The Δ is the spacing of ordered uniform distribution, which is $f_\Delta(\Delta) = N(1-\Delta)^{N-1}$ distributed.[2] Rewriting Equation (S18) in term of Δ yield

$$d_{DN} = \sqrt{\frac{\sum_{i=1}^{N}\left(\Delta_i(\frac{1}{N\Delta_i}-1)\right)^2}{\sum_{i=1}^{N}\Delta_i^2}}, \tag{S21}$$

in the limit where N is large ($N > 50$), we can apply the continuous approximation,

$$d_{DN}^{(*)} = \sqrt{\frac{\int_0^1 \left(\Delta(\frac{1}{N\Delta}-1)\right)^2 N(1-\Delta)^{N-1}\,d\Delta}{\int_0^1 \Delta^2 N(1-\Delta)^{N-1}\,d\Delta}} = \sqrt{\frac{1}{2}+\frac{2-N}{2N^2}}. \tag{S22}$$

where the $(*)$ indicates the analytically derived distribution noise, we have performed the integration in the surd, using the properties of the Beta distribution. Equation (S22) shows $d_{DN}^{(*)} \to 1/\sqrt{2}$ as $N \to \infty$ as illustrated in Figure S3.

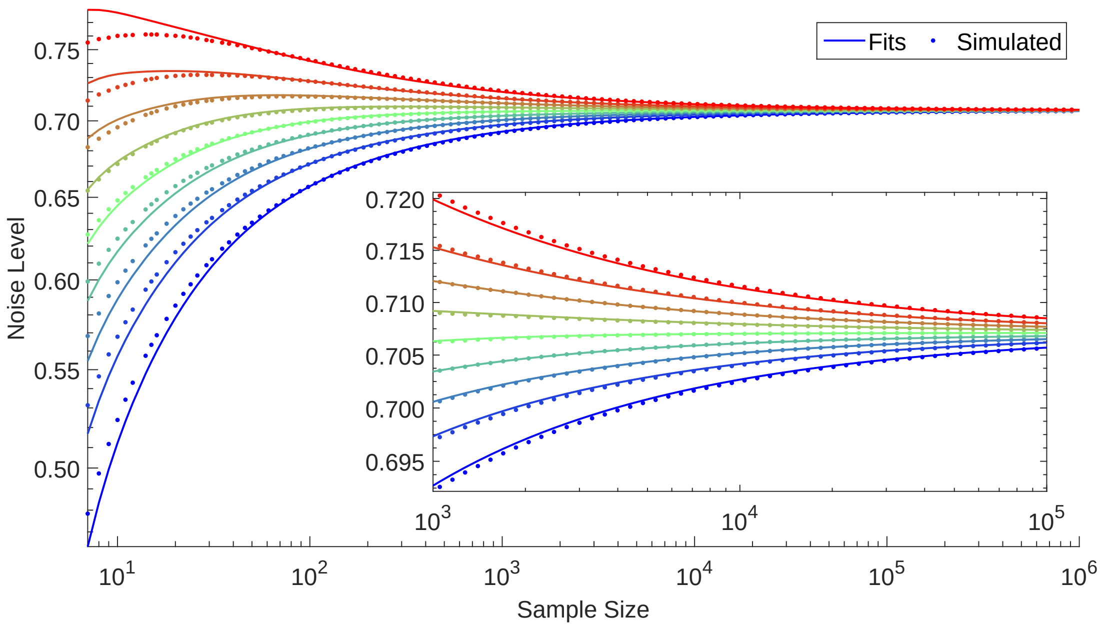

Figure S3: Relationship between distribution noise d_{DN} and sample size N at deciles going from the 10^{th} percentile (blue) to the 90^{th} (red), obtained from 10^6 simulations. The d_{DN} value converges to $1/\sqrt{2}$ as N increases.

2.2 Distribution Noise and Sample Size

We show the absolute deviation of d_{DN} value for the expected value $d_{DN}^{(*)}$ at different deciles Figure S4. Again there are straight lines on log-log plot, which indicate the power law relationship as $d_{DN} - \langle d_{DN}\rangle \propto A_{DN}N^{-\alpha_{DN}}$ for $N > 50$.

Thus, by assuming the proportionality constant A_{DN} depends only the percentile level $\wp_{DN}$, after experimenting with several functional forms, we write down the relationship between d_{DN} and N at

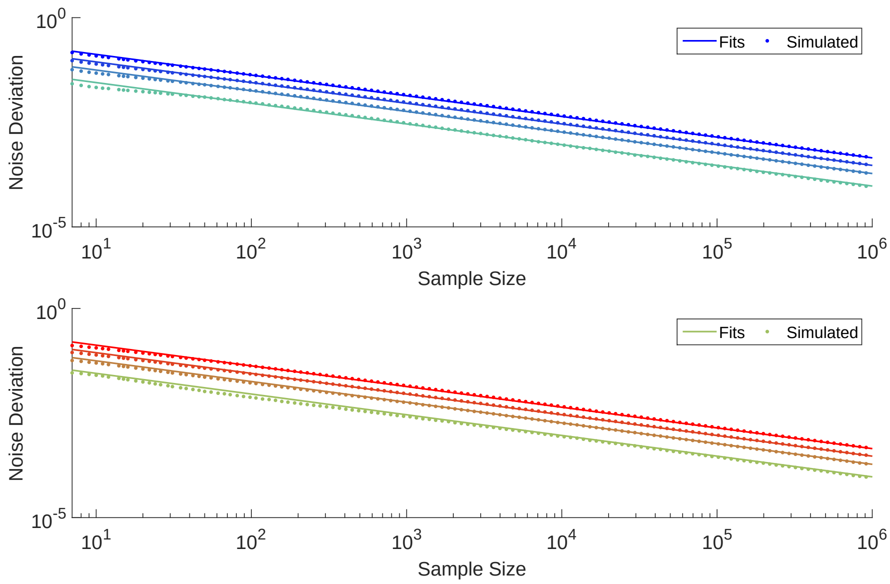

Figure S4: Relationship between absolute deviation of distribution noise from expected value, $|(d_{DN} - d_{DN}^{(*)}|$ and sample size N at various deciles obtained from 10^6 simulations. The top figure shows deciles from the 10^{th} percentile (blue) to to the 40^{th} (green), while the bottom shows decile from the 60^{th} (green) to the 90^{th} (red).

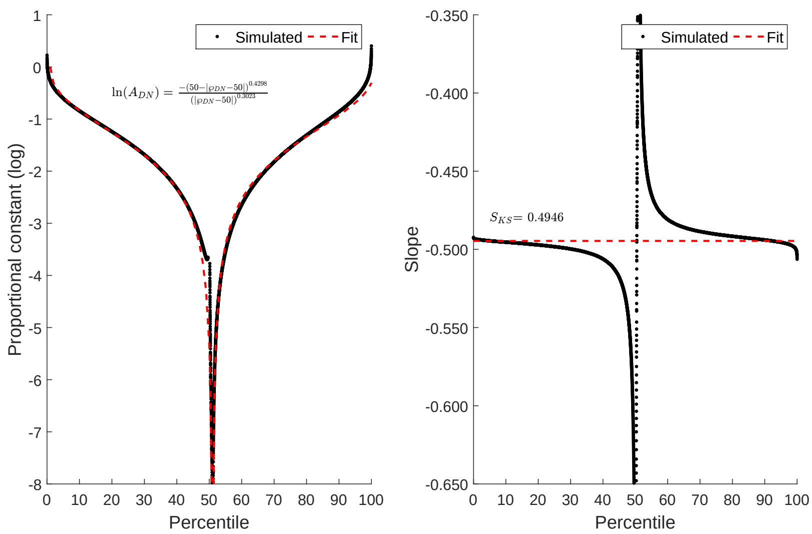

Figure S5: Fits for A_{DN} and α_{DN} at different percentile.

percentile $\wp_{DN}$ as

$$A_{DN} = \Phi(\wp_{DN} - 50) \exp\left(-\frac{[50 - |\wp_{DN} - 50|]^{0.430}}{|\wp_{DN} - 50|^{0.302}} \right), \tag{S23}$$

where $\Phi(x)$ represents the sign of x. For simplicity we assume constant power law exponent $\alpha_{DN} = 0.495$, which is average of the slopes in log-log plot excluding $\wp_{DN} \in [40, 60]$. Refer to Figure S5 for the fits of proportionality constant and power law exponent at different percentile. Hence, the equation that relate d_{DN} and N at percentile $\wp_{DN}$ is

$$d_{DN}(\wp_{DN}, N) = \langle d_{DN}\rangle + \Phi(\wp_{DN} - 50)\frac{\exp\left(-\frac{[50-|\wp_{DN}-50|]^{0.430}}{|\wp_{DN}-50|^{0.302}}\right)}{N^{0.495}}, \tag{S24}$$

where $\Phi(x)$ represents the sign of x, and

$$\langle d_{DN}\rangle = \sqrt{\frac{1}{2} + \frac{2-N}{2N^2}}\left(\frac{N}{N+0.5}\right) \tag{S25}$$

is the mean of d_{DN} that converges to $1/\sqrt{2}$ as $N \to \infty$, and $N/(N + 0.5)$ is added in as a correction factor for small N. Again this result also suggests that if we have two samples with sizes N_1 and N_2 with $N_2 > N_1$ from the same distribution, we should compare $N_1^{0.495}(d_{DN}^{(1)} - 1/\sqrt{2})$ against $N_2^{0.495}(d_{DN}^{(2)} - 1/\sqrt{2})$. Otherwise, we will make wrong conclusion that the N_2 sample fits the distribution better.

2.3 Correlation Between KS Distance and Distribution Noise

Since d_{KS} and d_{DN} are both measure of deviations, we need to account for the inherent correlation between these two variables, when applying them simultaneously to significance testing. To do this, we compute the Pearson correlation between d_{KS} and d_{DN} obtained from the 10^6 simulated samples for various sample sizes. Figure S6 shows the fits of the correlation coefficients,

$$\rho_{d_{KS},d_{DN}}(N) = \frac{e}{N^{0.481}}. \tag{S26}$$

and confirms our intuition that d_{KS} is positively correlated with d_{DN}. Since d_{KS} is a measure at the cumulative level, the correlation between d_{KS} and d_{DN} vanishes as $N \to \infty$, since the distribution noises cancel each other in this limit.

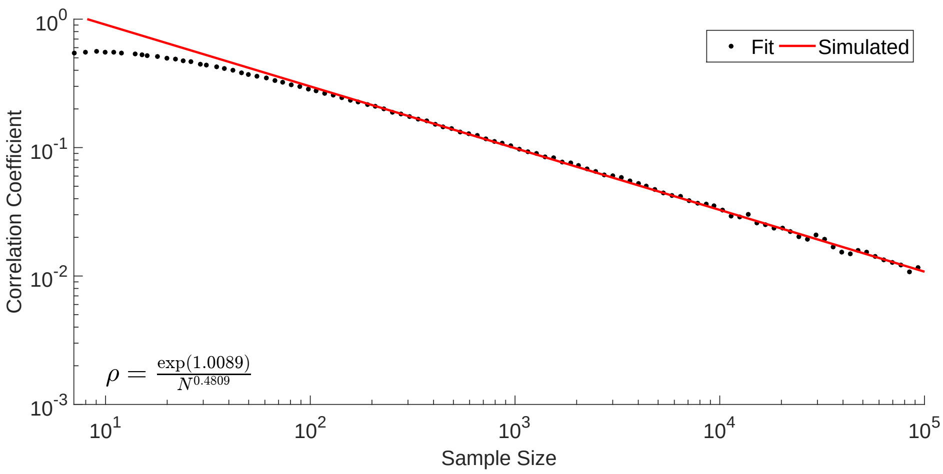

Figure S6: Correlation between d_{KS} and d_{DN} at different sample sizes, obtained from the 10^6 simulated samples.

3 REAL DATA DESCRIPTION

3.1 Taiwan Housing Data

The Taiwanese housing price shown in text is obtained from the Department of Land Administration (`http://plvr.land.moi.gov.tw/DownloadOpenData`). All housing transaction from 2010 are available. Following a previous study,[3] we used only data from the Greater Taipei Area, containing Taipei city, New Taipei city, Keelung City and Taoyuan city, from 2010 to 2012 for the Zhu Zai Da Lous (which are condominium type housing). We understand from Ref. [3] that the distribution contains statistically significant outliers at the tail and choose to truncate the data set at 4.5×10^5, which is the $x_{min}^{(PL)}$, to avoid the scenario where the distribution is polluted by outliers.

3.2 Taiwan Income Data

The Taiwan income data set contains information on the survey of household income in Taiwan. The summary of its findings can be found at `http://win.dgbas.gov.tw/fies/index.asp` while the raw data set is available through a paid subscription which can be obtained by completing a form at `http://win.dgbas.gov.tw/fies/order.asp`. Following a previous studies reviewed in Ref. [4], the income distribution generally has an exponential body and power-law tail. We use the Clauset-Shalizi-Newman method for PL [5] to obtain the x_{min} of the PL part of the Taiwan Income distribution. Next, we to truncate the full data set at this value $x_{min}^{(PL)}$, which leaves us with only the exponential part of the distribution. In the main text, we work with data containing on the EXP part of the distribution.

3.3 SGX Normalized Returns

It is well known from previous works that the short interval (~ 5 mins) normalized returns of stock price indices follow a PL distribution.[6, 7, 8] We check this result on the the Straits Times Index (STI), which tracks the Singapore Stock Exchange (SGX), using the methods discussed in this paper. Since most of the public sources only provide historical data for the daily close price, we need to constructed the index using the tick-historical data in order to get fine-grained, short-interval (~ 5 mins) data. To construct the STI price time series, we choose the price for a particular interval by considering the tick price that is closest the end of the 5-min interval. We do this for all component stocks of the STI during the period 2009 to 2016. We note that our index price time series is not entirely similar to the published STI, since the STI is capital weighted. Since we do not have the capital information of the underlying company, we used a simple average weighting formula to construct the index. However, we do not expect this difference in weighting to change the PL characteristics of its return distribution, as many of the stock indices returns (cited above) is PL distributed but generally have different weighting formulas.

Next we obtain the set of returns $\mathbf{r} = r_1, r_2, \ldots, r_N$, by calculating the returns in each 5-min interval using:

$$r_t = \frac{p_t - p_{t-1}}{p_{t-1}}. \tag{S27}$$

Where r_t and p_t are the return and index price at interval t, and p_{t-1} is index price of the previous $(t-1)$ interval. The normalized return $\mathbf{r}^{(n)}$ is then calculated using

$$r_t^{(n)} = \frac{r_t - \langle \mathbf{r} \rangle}{\sigma(\mathbf{r})}, \tag{S28}$$

where $\langle \mathbf{r} \rangle$ and $\sigma(\mathbf{r})$ are respectively, the mean and standard deviation of $\mathbf{r}$.

4 FITS TO REAL DATA

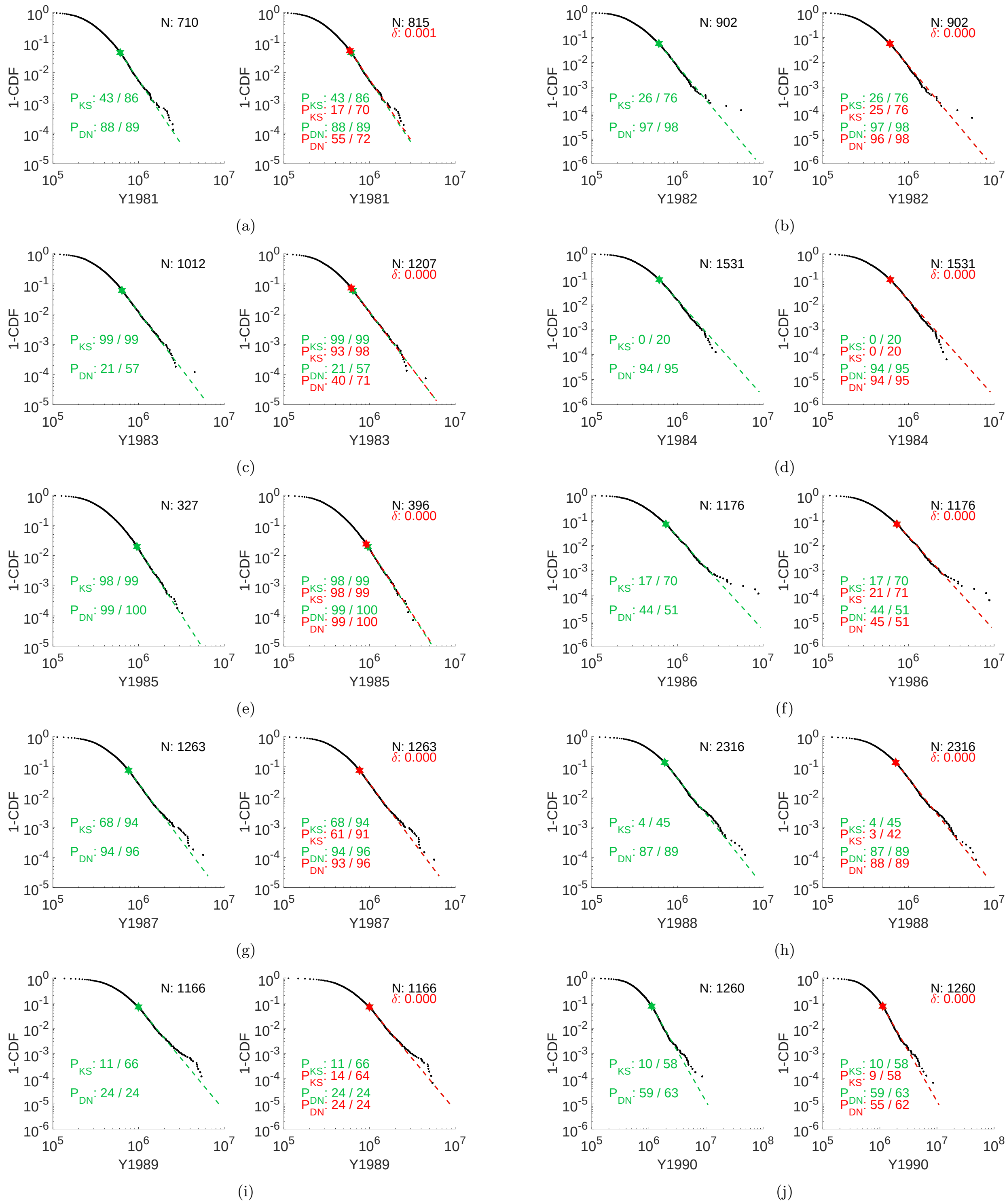

Figure S7: Year by year Taiwan income data from 1981 to 1990, fitted to power law distributions. In all figures, the empirical CDF that is adjusted for FLE is shown as black dots, while the unadjusted and adjusted fits are shown as green and red dashed line respectively. $P_{KS/DN}$-values (in percentage) are for unadjusted (green) and adjusted (red) fits. We separate the p-values obtained using the CSN method (left) from those using inversion formulae (right) by a '/'.

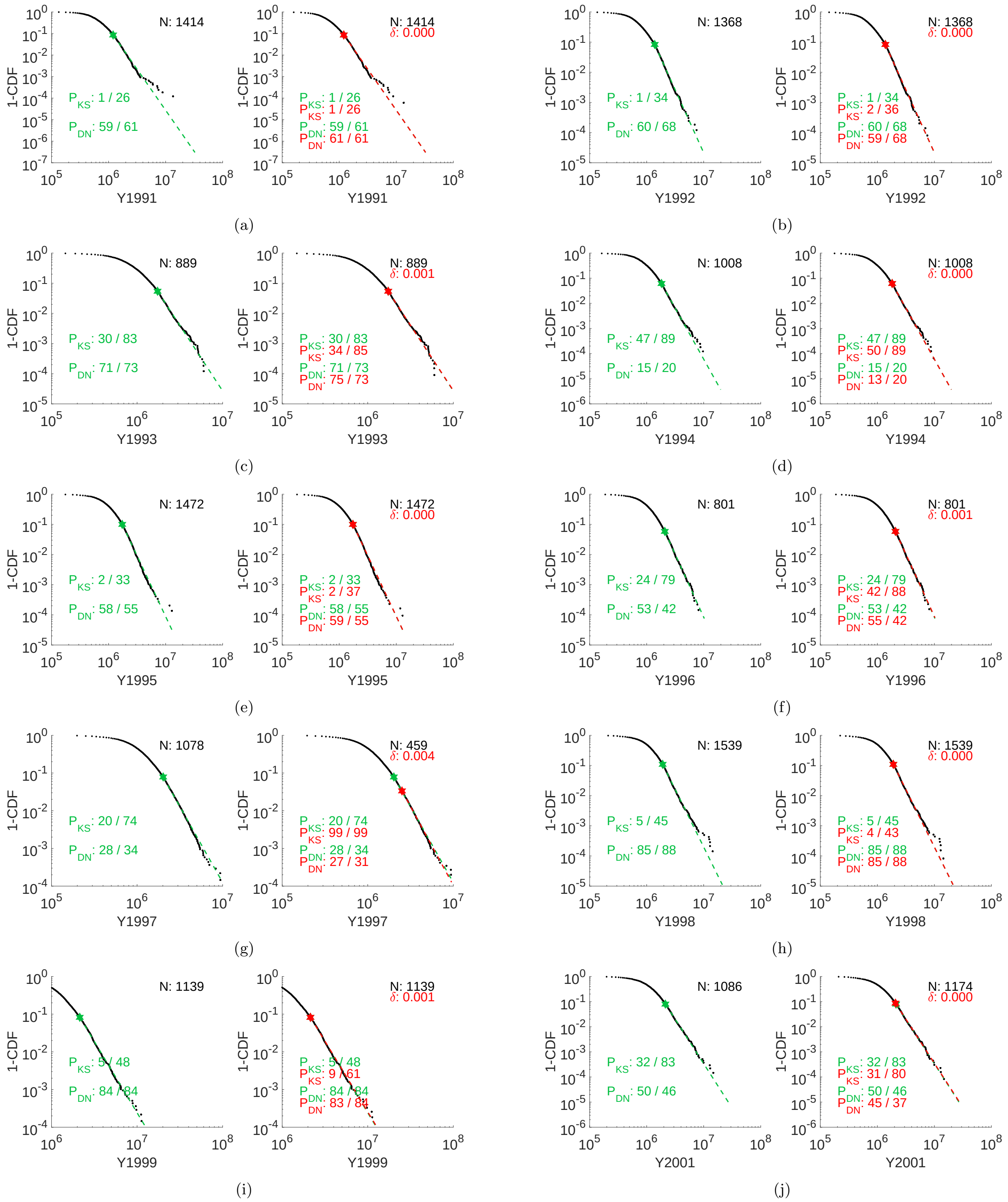

Figure S8: Year by year Taiwan income data from 1991 to 2001, fitted to power law distributions. In all figures, the empirical CDF that is adjusted for FLE is shown as black dots, while the unadjusted and adjusted fits are shown as green and red dashed line respectively. $P_{KS/DN}$-values (in percentage) are for unadjusted (green) and adjusted (red) fits. We separate the p-values obtained using the CSN method (left) from those using inversion formulae (right) by a '/'.

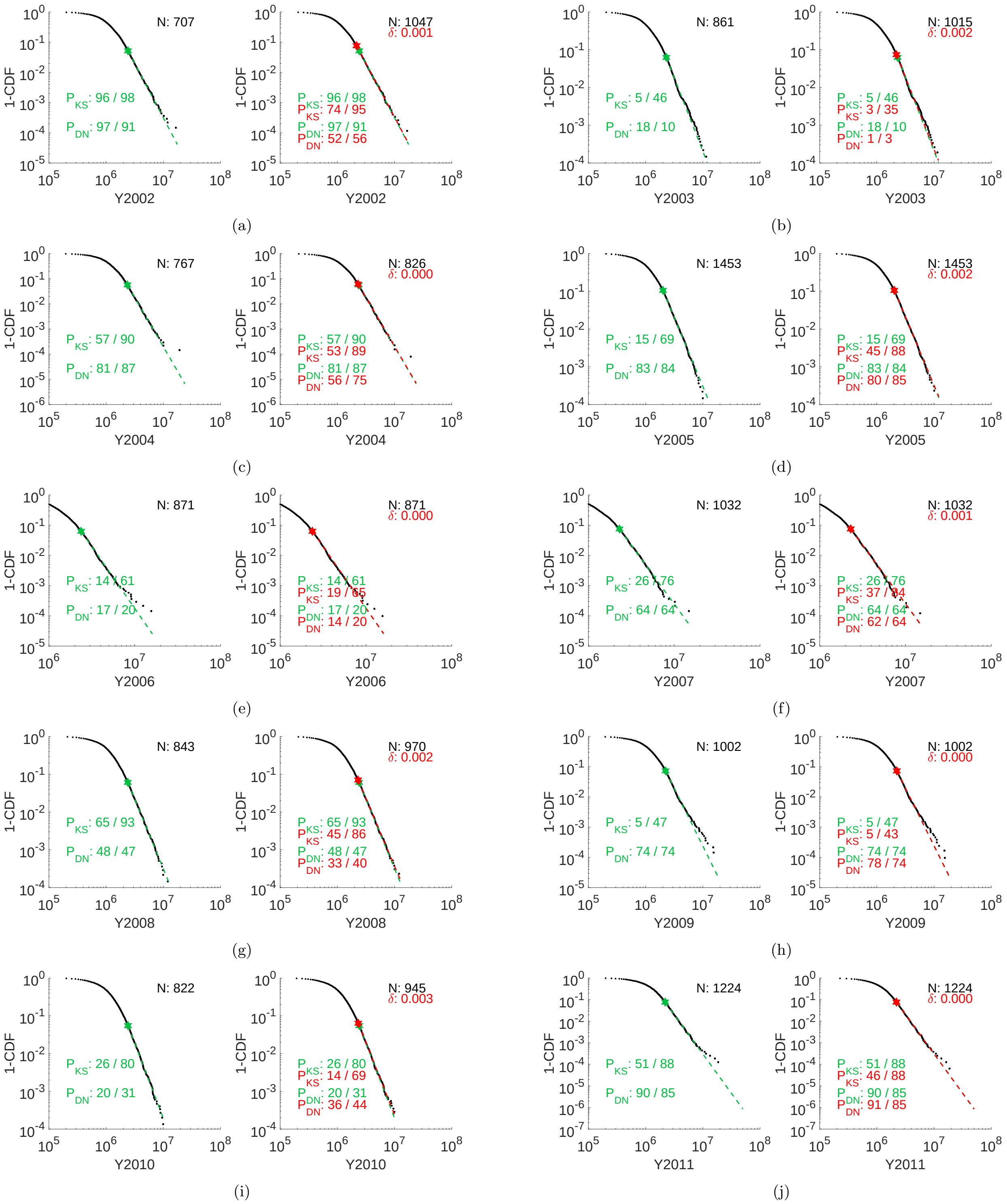

Figure S9: Year by year Taiwan income data from 2002 to 2011, fitted to power law distributions. In all figures, the empirical CDF that is adjusted for FLE is shown as black dots, while the unadjusted and adjusted fits are shown as green and red dashed line respectively. $P_{KS/DN}$-values (in percentage) are for unadjusted (green) and adjusted (red) fits. We separate the p-values obtained using the CSN method (left) from those using inversion formulae (right) by a '/'.

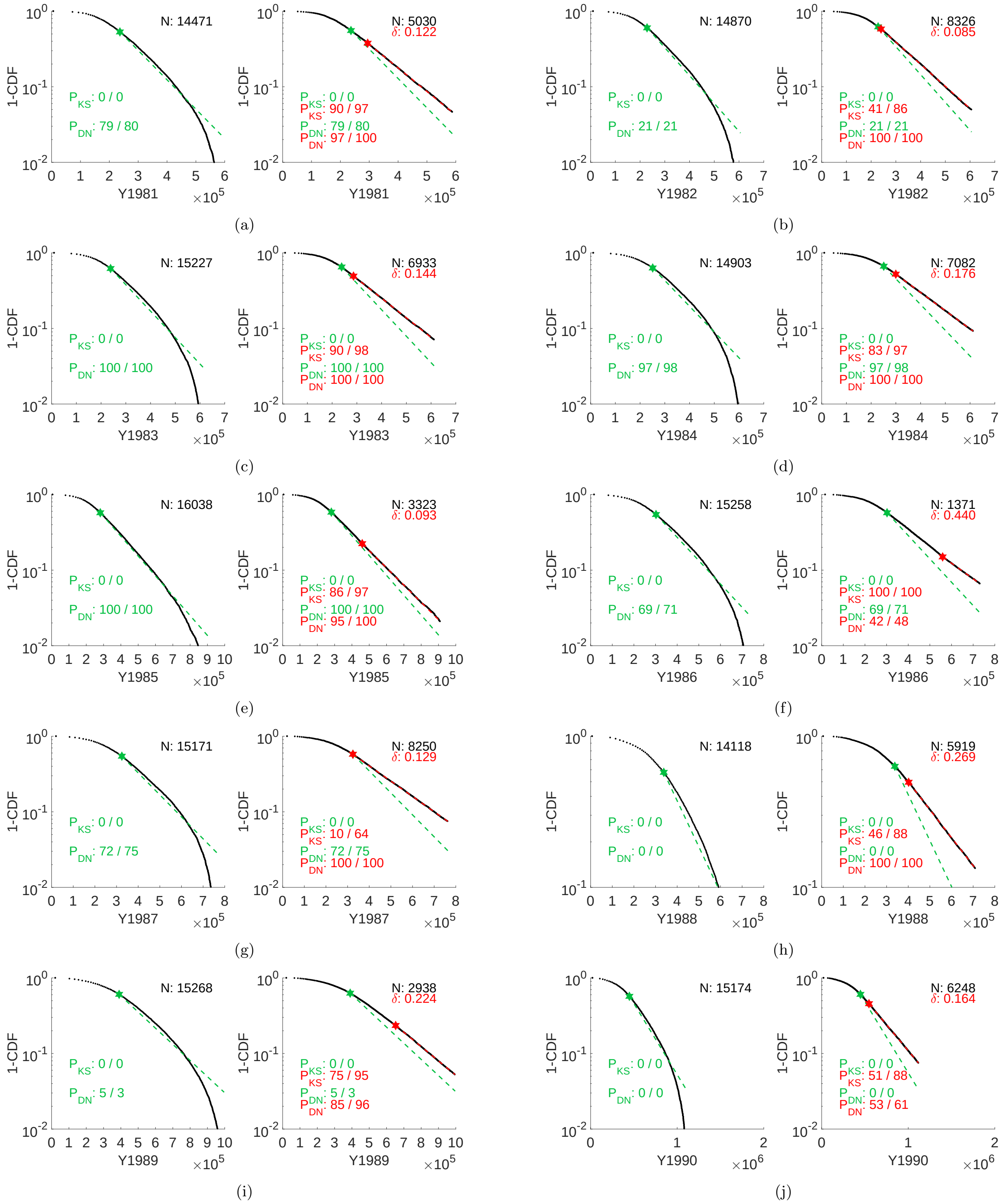

Figure S10: Year by year Taiwan income data from 1981 to 1990, fitted to exponential distributions with $x_{max}^{(EXP)} = x_{min}^{(PL)}$. In all figures, the empirical CDF that is adjusted for FLE is shown as black dots, while the unadjusted and adjusted fits are shown as green and red dashed line respectively. $P_{KS/DN}$-values (in percentage) are for unadjusted (green) and adjusted (red) fits. We separate the p-values obtained using the CSN method (left) from those using inversion formulae (right) by a '/'.

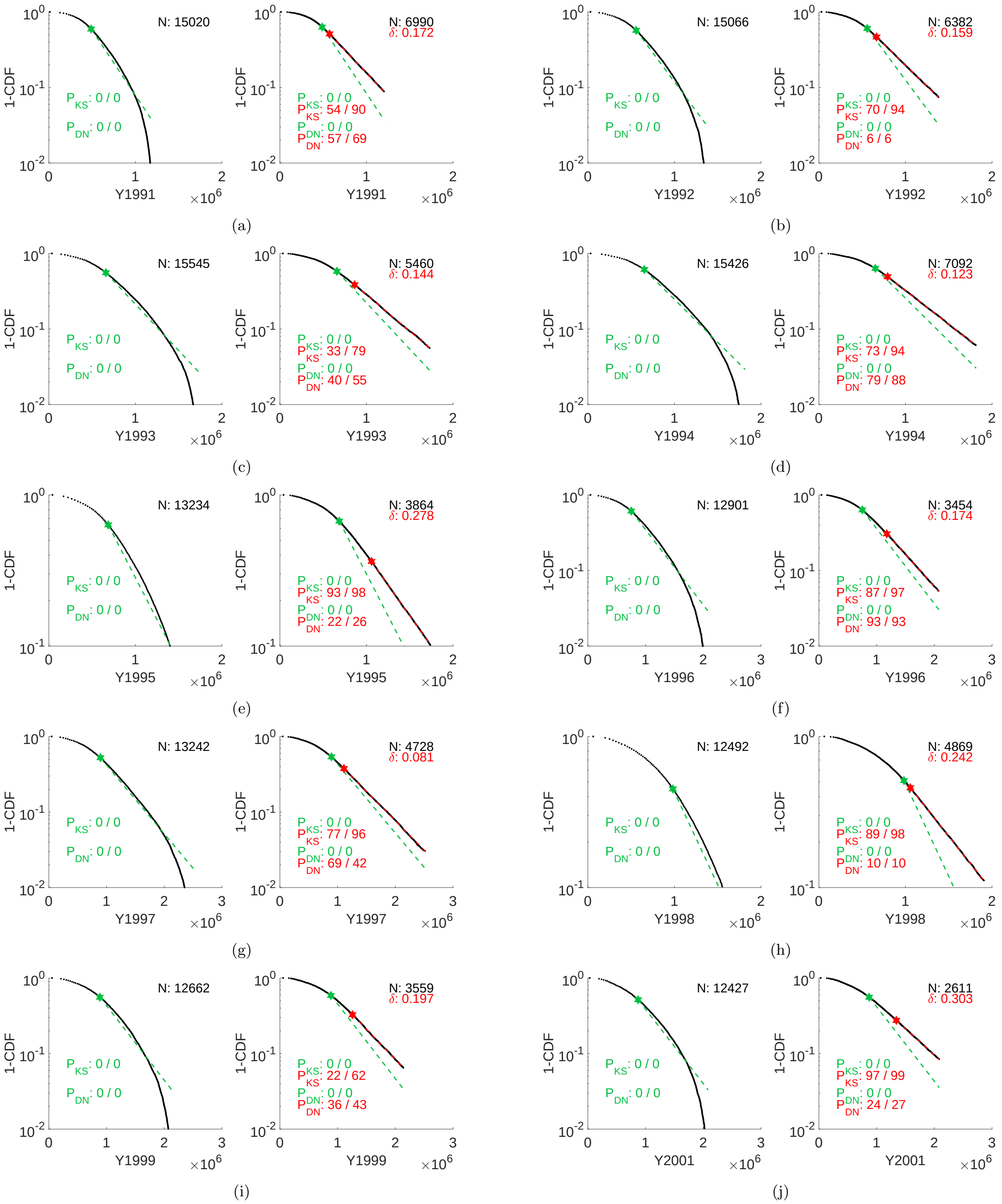

Figure S11: Year by year Taiwan income data from 1991 to 2001, fitted to exponential distributions with $x_{max}^{(EXP)} = x_{min}^{(PL)}$. In all figures, the empirical CDF that is adjusted for FLE is shown as black dots, while the unadjusted and adjusted fits are shown as green and red dashed line respectively. $P_{KS/DN}$-values (in percentage) are for unadjusted (green) and adjusted (red) fits. We separate the p-values obtained using the CSN method (left) from those using inversion formulae (right) by a '/'.

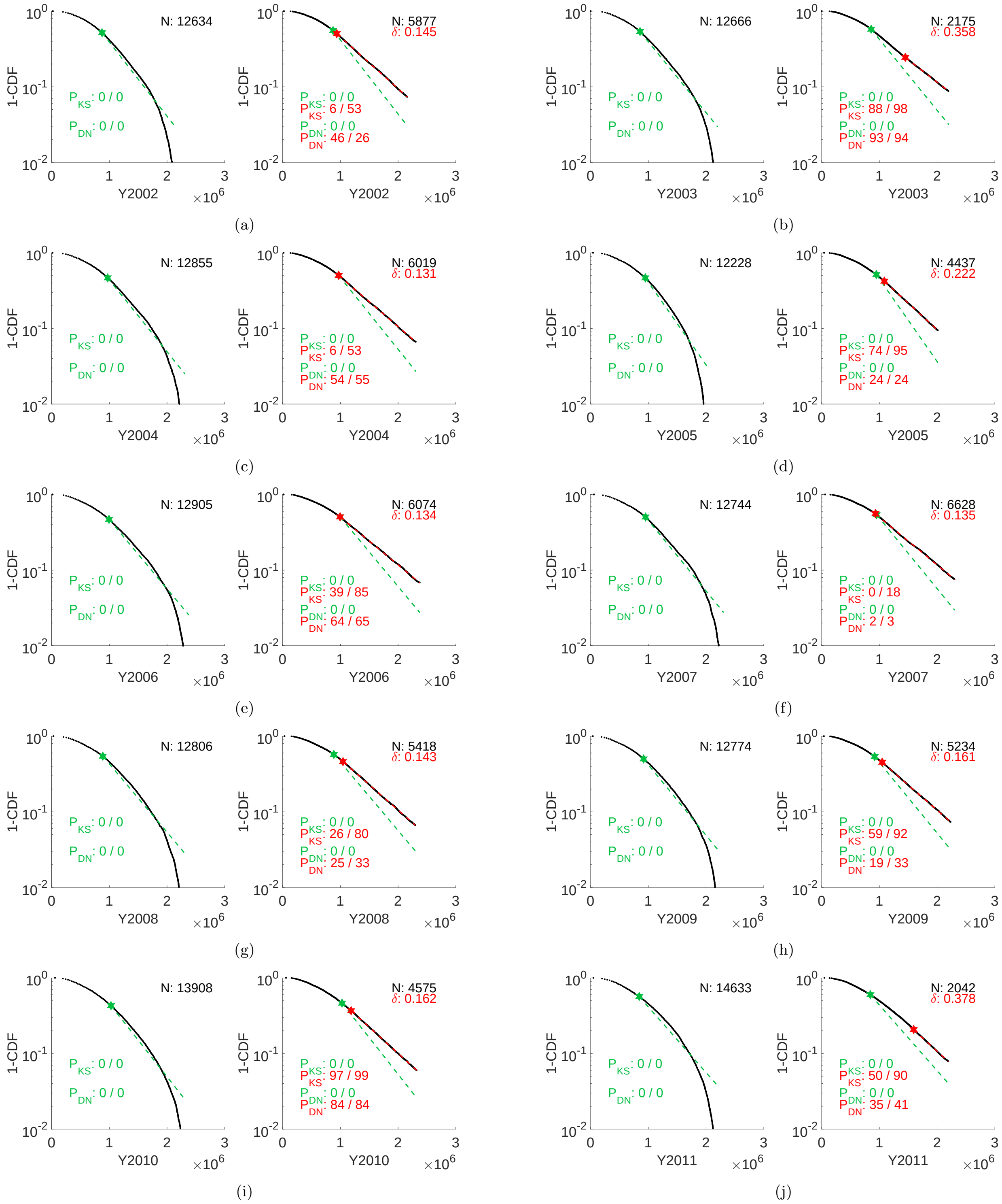

Figure S12: Year by year Taiwan income data from 2002 to 2011, fitted to exponential distributions with $x_{max}^{(EXP)} = x_{min}^{(PL)}$. In all figures, the empirical CDF that is adjusted for FLE is shown as black dots, while the unadjusted and adjusted fits are shown as green and red dashed line respectively. $P_{KS/DN}$-values (in percentage) are for unadjusted (green) and adjusted (red) fits. We separate the p-values obtained using the CSN method (left) from those using inversion formulae (right) by a '/'.

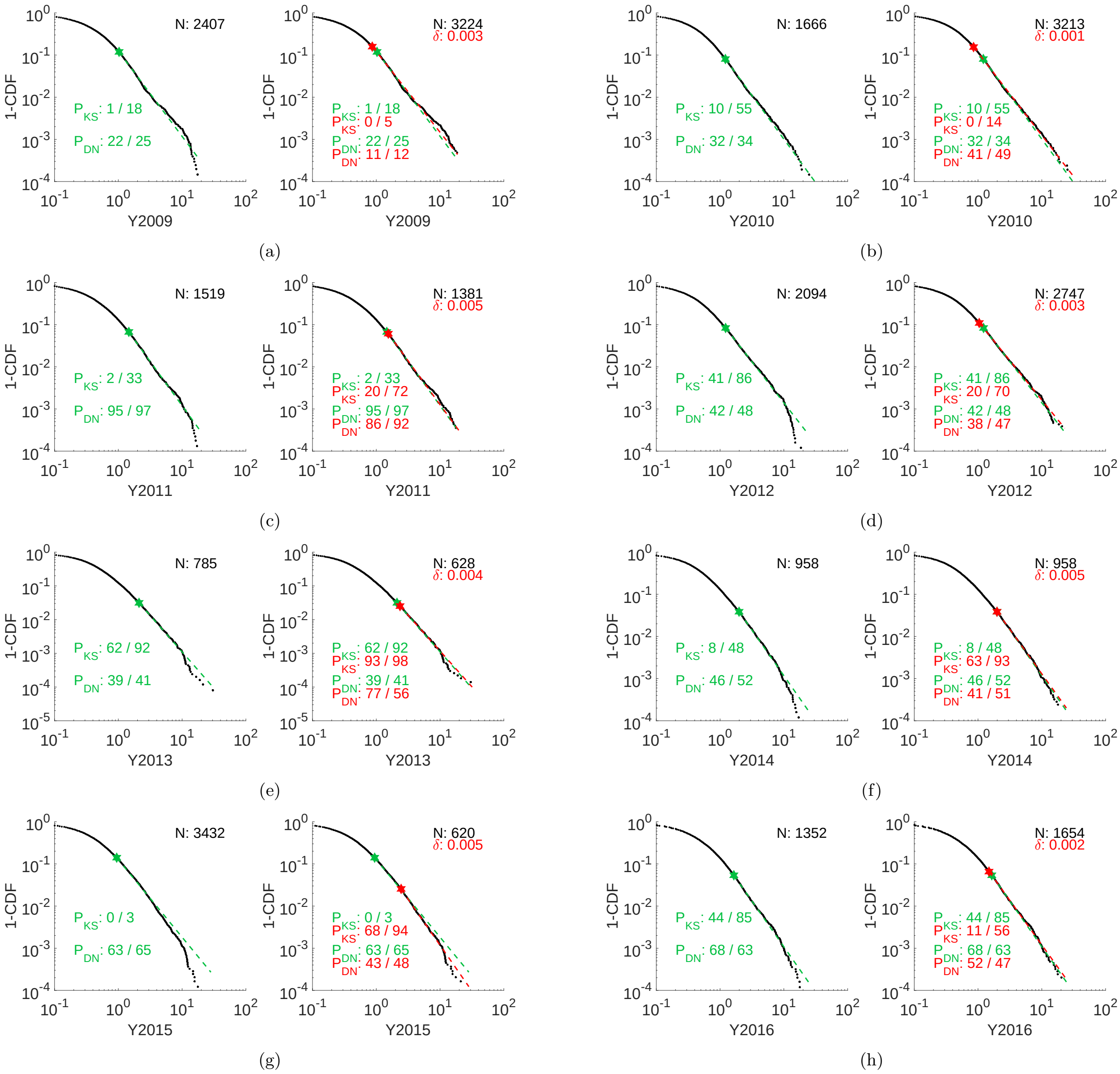

Figure S13: Year by year SGX normalized returns from 2009 to 2016, fitted to power law distributions. In all figures, the empirical CDF that is adjusted for FLE is shown as black dots, while the unadjusted and adjusted fits are shown as green and red dashed line respectively. $P_{KS/DN}$-values (in percentage) are for unadjusted (green) and adjusted (red) fits. We separate the p-values obtained using the CSN method (left) from those using inversion formulae (right) by a '/'.

REFERENCES

[1]Luc Devroye. *Non-uniform random variate generation*. Springer, 1 edition, 1986.
[2]Ronald Pyke. Spacings. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 395–449, 1965.
[3]Darrell Jiajie Tay, Chung-I Chou, Sai-Ping Li, Shang You Tee, and Siew Ann Cheong. Bubbles are departures from equilibrium housing markets: Evidence from singapore and taiwan. *PloS one*, 11(11):e0166004, 2016.
[4]Victor M Yakovenko and J Barkley Rosser Jr. Colloquium: Statistical mechanics of money, wealth, and income. *Reviews of Modern Physics*, 81(4):1703, 2009.
[5]Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.
[6]Rosario N Mantegna and H Eugene Stanley. Scaling behaviour in the dynamics of an economic index. *Nature*, 376(6535):46–49, 1995.
[7]Parameswaran Gopikrishnan, Vasiliki Plerou, Luis A Nunes Amaral, Martin Meyer, and H Eugene Stanley. Scaling of the distribution of fluctuations of financial market indices. *Physical Review E*, 60(5):5305, 1999.
[8]Boon Kin Teh and Siew Ann Cheong. The asian correction can be quantitatively forecasted using a statistical model of fusion-fission processes. *PloS one*, 11(10):e0163842, 2016.