Supplementary Information
Artificial Increasing Returns to Scale and the Problem of Sampling from Lognormals

Supplementary Information A. Validating the assumption that the sum $Y(n)$ can be approximated by the maximum $M(n)$

We approximate $Y(n)$ by the maximal productivity $M(n) := \max\{X_1, \dots, X_n\}$ in the city. This quantity can be written as $M(n) = e^{\sigma L(n) - \sigma^2/2}$, where $L(n) := \max\{Z_1, \dots, Z_n\}$ denotes the maximum of i.i.d. standard normal random variables. Then

$$\begin{aligned} Y(n) &= \sum_{i=1}^{n} X_i \\ &= \sum_{i=1}^{n} \mathrm{e}^{\sigma Z_i - \sigma^2/2} \\ &= \mathrm{e}^{\sigma L(n) - \sigma^2/2} \sum_{i=1}^{n} \mathrm{e}^{\sigma (Z_i - L(n))} \\ &= M(n) \sum_{i=1}^{n} \mathrm{e}^{\sigma (Z_i - L(n))}. \end{aligned} \tag{16}$$

The main difficulty for validating the assumption that $Y(n)$ can be approximated by $M(n)$ is in analyzing the last sum in Equation (16). Since it is doubtful that this quantity can be tackled analytically, we suggest the following argument. First write

$$\Delta_n := \sum_{i=1}^{n} e^{\sigma(Z_i - L(n))} = \sum_{i=1}^{n} \mathrm{e}^{\sigma(L_i(n) - L(n))},$$

where we have re-ordered the terms in the summation such that $L_i(n)$ denotes the ith largest value among $Z_1, \dots, Z_n$. For the first term, we have $L_1(n) = L(n)$, so $\mathrm{e}^{\sigma(L_1(n) - L(n))} = 1$. For the second term, we can use that $L(n) - L_2(n)$ is of order $(\ln(n))^{-1/2}$ (see Leadbetter et al. 1983, Section 2.3). By our assumption that $\sigma \gg \sqrt{2\ln(n)}$, the quantity $\sigma(L_2(n) - L(n))$ is negatively large and so $\mathrm{e}^{\sigma(L_2(n) - L(n))}$ is close to 0. The remaining terms $\mathrm{e}^{\sigma(L_i(n) - L(n))}$ for $i \geq 3$ decay to 0 much faster since so do exponents with larger negative powers.

Thus, we have $\Delta_n \approx 1$ when $\sigma \gg \sqrt{\ln(n)}$. Moreover, our simulations (not shown) reveal that a similar conclusion applies even when σ is larger than, but comparable to, $\sqrt{2\ln(n)}$, in which case Δ_n is rather close to 1, being of constant order.

Figure 6 illustrates the fact that the maximum can indeed become comparable to the sum. We use a proxy of the share $M(n)/Y(n)$ as the ratio of the quantile $Q(1 - 1/n)$ over $\sum_i Q(i/n - 1/n)$, where $Q(\Pr(X \leq x_p)) = x_p$. The figure shows the curves for this proxy of the share $M(n)/Y(n)$ as a function of σ, for four distinct values of n.

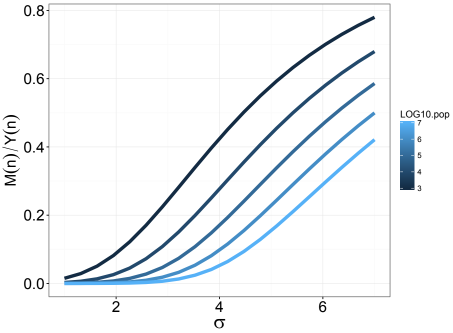


Figure 6. Proxy for the share of the maximum over the total sum, $M(n)/Y(n)$, constructed as the ratio of the quantile of the lognormal associated with the percentile $(n-1)/n$ over the sum of all the n quantiles, as a function of parameter σ. The color of the line represents a fixed population size. We show the curves for $n = 10^3$, 10^4, 10^5, 10^6, and 10^7. The lighter the blue is, the larger the population is.

According to our derivations, for $\sigma = 4$, the maximum is comparable to the sum when $n < \mathrm{e}^{\sigma^2/2} \approx 3,000$. Indeed, the figure shows that for $\sigma = 4$ and for $n = 10^3$ (darkest blue line), the maximum can account for 50% of the sum. For $\sigma = 4$ one needs to increase size to $n = 10^7$ (a ten-thousand-fold increase) in order to decrease the dominance of the maximum to about 10% (see lightest blue line). Figure 6 provides evidence in support of replacing the sum $Y(n)$ with the maximum $M(n)$, and using this to find an approximate result for how the sum scales with size.

Supplementary Information B. Data

In the main text we use data of the formal workforce in Colombia to analyze the unconditional elasticity of nominal wages on municipality population size. These come from the administrative records of the social security system in Colombia (abbreviated as PILA in Spanish, meaning the *Integrated Report of Social Security Contributions*). The PILA is maintained by the Colombia Ministry of Finance and Public Credit ("Ministerio de Hacienda y Crédito Público"). PILA consists of individual contributions to health and pensions reported by workers, firms, public institutions, and other formal entities like associations, universities, cooperatives and multilateral organizations.

The dataset was obtained from the Ministry of Finance and Public Credit, under a data use agreement that is part of the development of `www.datlascolombia.com`, a joint project between the Center for International Development and the Colombian Foreign Trade Bank (Bancoldex) to map the industrial economic activity in Colombia. The data are stored on secure computers at the Harvard-MIT Data Center. Access is restricted to identified and authorized researchers by means of a confidential account. The use of the PILA for research purposes has been reviewed by the Harvard's Institutional Review Board (IRB). In the database individuals and firms have been previously anonymized in order to protect their habeas data. Harvard IRB determined that this dataset is not human subjects as defined by the Department of Health and Human Services (DHHS) regulations.

Each row of the dataset consists of a monthly contribution to the social security system, with more than seventy different fields with information about the worker and the firm, and with the values of the contribution to health and pension, according to the days the worker worked at the firm in that month. The raw microdata consists of 122,287,562 rows (i.e., social security contributions), from 10,535,587 unique workers (i.e., each worker had an average of 11.6 contributions per year). As explained below, we aggregate and keep a subset of all these observations, and we only use two fields for this study: the list of nominal wages earned, and the municipalities of work to which the wage values where attached.

As a start, these data must be cleaned, as is often the case with datasets built from observations resulting from administrative transactions. Common problems include misreported or missing wages, no municipality of work reported, no age reported, duplicated observations, or missing contribution to pension or health. In addition to dropping these problematic observations, we keep only those workers that are categorized as "dependent" or "independent", which means they are either employed in a firm or are self-employed, respectively (by keeping these type of social security contributors we exclude those individuals that contribute to social security through means other than a formal job). Finally, we keep those individuals who worked for at least 30 days during the whole year, and had ages between 15 and 64.

We compute the monthly average wage of workers by first adding their net wage earned during the year, then dividing it by the total number of days worked, and finally multiplying by thirty. By law, firms are required to pay a minimum wage to workers, or more. However, there exist special cases in the dataset in which this does not hold. Hence, we make sure this is the case by dropping observations which report average monthly wages below the minimum wage ($616,000 Colombian Pesos, or COP, in 2014). At the end, our population of analysis consists of 6,713,975 formal Colombian workers (approximately 64% of the unique individuals that appear originally in the dataset).

Supplementary Information C. Derivation of $\beta_{ave}(n_{\min}, \sigma, \alpha)$

Here we shall indicate the relevant steps for the derivation of Equation (12).

The derivation becomes relatively easy once some changes of variables are carried out first. We start with the change of variable $U = \alpha \ln(N)$. Together with equation (10) where $N \sim Pareto(n_{\min}, \alpha)$, and the conservation of probability, we get that $p(u) = p(n) \left|\frac{\mathrm{d}u}{\mathrm{d}n}\right|^{-1} = \mathrm{e}^{-(u-q)}$, for $u \geq q$, where $q \equiv \alpha \ln(n_{\min})$. That is, a shifted standard exponential,

$$U \sim q + Exp(1).$$

With the additional change of variable $V = \ln(Y)$, the piecewise function in equation (9) can be re-written as

$$\mathrm{E}\left[V|U\right] = \begin{cases} \frac{\sigma^2 U}{2w} & \text{, for } U \geq w \\ -\frac{\sigma^2}{2} + \frac{\sigma^2}{\sqrt{w}}\sqrt{U} & \text{, for } U < w, \end{cases}$$

where $w \equiv \alpha\sigma^2/2$. Using the indicator function, we express the equation above more concisely as

$$\mathrm{E}\left[V|U\right] = \frac{\sigma^2}{2w}\left[U + \left(2w^{1/2}U^{1/2} - w - U\right)\mathbb{1}_{\{U<w\}}\right].$$

For the computation of $\beta_{ave}(n_{\min}, \sigma, \alpha)$ we have to get analytic expressions of the different terms in the following relation (see main text):

$$\begin{aligned}\beta_{ave}(n_{\min}, \sigma, \alpha) &= \frac{\mathrm{E}\left[(U/\alpha)V\right] - \mathrm{E}\left[(U/\alpha)\right]\mathrm{E}\left[V\right]}{\mathrm{Var}\left[(U/\alpha)\right]} \\ &= \alpha\left(\frac{\mathrm{E}\left[UV\right] - \mathrm{E}\left[U\right]\mathrm{E}\left[V\right]}{\mathrm{Var}\left[U\right]}\right).\end{aligned}$$

The easy terms are those in which U is alone:

$$\begin{aligned}\mathrm{E}\left[U\right] &= 1 + q, \\ \mathrm{Var}\left[U\right] &= 1.\end{aligned}$$

The piecewise form of $\mathrm{E}\left[V|U\right]$, however, complicates the rest. Let us compute $\mathrm{E}\left[V\right]$ using the law of total expectations $\mathrm{E}\left[V\right] = \mathrm{E}\left[\mathrm{E}\left[V|U\right]\right]$:

$$\mathrm{E}\left[V\right] = \int_q^\infty \frac{\sigma^2}{2w}\left[u + \left(2w^{1/2}u^{1/2} - w - u\right)\mathbb{1}_{\{u<w\}}\right]p(u)\mathrm{d}u.$$

The first term is trivial, since it is simply the expectation of U. For the second term, since the integral starts at q, the indicator function $\mathbb{1}_{\{u<w\}}$ can only be true whenever w is also larger than q. Hence, we can compute the second term in the integral by taking the indicator function out of the integral and replacing it with $\mathbb{1}_{\{q<w\}}$, and evaluating the integral only between q and w:

$$\mathrm{E}\left[V\right] = \frac{\sigma^2}{2w}\left[(1+q) + \mathbb{1}_{\{q<w\}}\int_q^w \left(2w^{1/2}u^{1/2} - w - u\right)p(u)\mathrm{d}u\right].$$

An almost identical expression can be derived for $\mathrm{E}\left[UV\right]$.

Now, given that $p(u) = \mathrm{e}^q\mathrm{e}^{-u}$, one recognizes that most terms become integrals of the form

$$\int_q^w u^s\mathrm{e}^{-u}\mathrm{d}u = \gamma(s+1, w) - \gamma(s+1, q),$$

for some constant s, where we use "lower incomplete gamma functions", defined as $\gamma(s+1, x) = \int_0^x u^s\mathrm{e}^{-u}\mathrm{d}u$. We use the recurrence relationship

$$\gamma(s+1, x) = s\gamma(s, x) - x^s\mathrm{e}^{-x}$$

iteratively, as many times as necessary until we either get to terms like $\gamma(1/2, x)$, or terms like $\gamma(1, x)$. For the former, we use the fact that $\sqrt{\pi}\mathrm{erf}(\sqrt{x}) = \gamma(1/2, x)$, to express everything in terms of “error functions”, defined as $\mathrm{erf}(x) = (2/\sqrt{\pi}) \int_0^x \mathrm{e}^{-t^2}\,\mathrm{d}t$. For the latter, we use the (trivial) fact that $\gamma(1, x) = 1 - \mathrm{e}^{-x}$.

After reducing the relevant integrals of $\mathrm{E}\left[V\right]$ and $\mathrm{E}\left[UV\right]$ to simple exponentials and error functions, it is just a matter of collecting terms. Finally, we arrive to the final expression:

$$\beta_{ave}(n_{\min}, \sigma, \alpha) = \begin{cases} 1, \quad \text{for } q \geq w \\ \\ \mathrm{e}^{q-w}(1-q) + (wq)^{1/2} \\ \qquad + \frac{\mathrm{e}^{q}(\pi w)^{1/2}}{2}(1-2q)\left(\mathrm{erf}(w^{1/2}) - \mathrm{erf}(q^{1/2})\right), \quad \text{for } q < w. \end{cases}$$

With some minor replacements, the reader can check that this is the same as equation (12) in the main text.

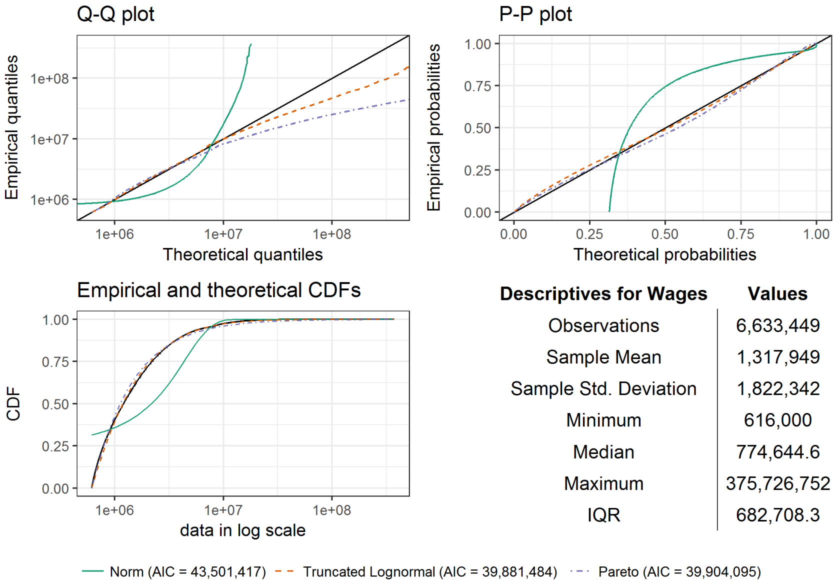
Supplementary Information D. Tables for goodness-of-fit statistics for monthly wages and municipality sizes in Colombia

Here, we show some tables comparing different alternative distributions and their goodness-of-fit for wages and sizes, and we show some comparative graphs.

Wages

Table 1. Distributions fitted to individual wages. The number of total workers ($1,325,950$ observations) analyzed in this table differ from the number mentioned in the main text ($6,633,449$) because wages are clustered on the minimum wage. The fits of continuous distributions to data with repeated values, such as the minimum value which is repeated several times, was much improved when we removed repeated values. The list of the distributions are ordered from top to bottom by increasing AIC values.

dist	numobs	loglik	AIC	BIC	Parameter 1	C.I.	Parameter 2	C.I.
trunclnorm	$1,325,950$	-$19,940,740$	$39,881,484$	$39,881,508$	$\widehat{\ln(x_0)} = 10.23$	[2, 10.15]	$\widehat{\sigma} = 2.00$	[1.99, 2.02]
powerlaw	$1,325,950$	-$19,952,047$	$39,904,095$	$39,904,108$	$\widehat{\alpha} = 1.16$	[1.16, 1.16]		
trunccauchy	$1,325,950$	-$19,960,472$	$39,920,948$	$39,920,972$	$\widehat{l} = 120,766.20$	[190129.96, 114817.91]	$\widehat{s} = 190,130.00$	[174740.73, 206786.87]
truncgamma	$1,325,950$	-$20,018,427$	$40,036,859$	$40,036,883$	$\widehat{a} = 0.0000$	[0, 0]	$\widehat{\lambda} = 0.0000$	[0, 0]
truncweibull	$1,325,950$	-$20,077,580$	$40,155,164$	$40,155,188$	$\widehat{a} = 0.72$	[0.72, 1109695.37]	$\widehat{b} = 1.11 \times 10^6$	[1109654.53, 1111878.12]
truncgumbel	$1,325,950$	-$20,330,914$	$40,661,832$	$40,661,856$	$\widehat{a} = 0.13$	[0.13, 1332540.63]	$\widehat{b} = 1.33 \times 10^6$	[1330795.1, 1332942.21]
lnorm	$1,325,950$	-$20,344,669$	$40,689,342$	$40,689,366$	$\widehat{\ln(x_0)} = 14.19$	[0.76, 14.19]	$\widehat{\sigma} = 0.76$	[0.76, 0.76]
gamma	$1,325,950$	-$20,626,561$	$41,253,126$	$41,253,151$	$\widehat{a} = 1.41$	[0, 1.41]	$\widehat{\lambda} = 0.0000$	[0, 0]
weibull	$1,325,950$	-$20,667,064$	$41,334,133$	$41,334,157$	$\widehat{a} = 1.05$	[1.05, 2219332.54]	$\widehat{b} = 2.22 \times 10^6$	[2219270.9, 2222105.32]
gumbel	$1,325,950$	-$20,825,629$	$41,651,261$	$41,651,285$	$\widehat{a} = 1.30 \times 10^6$	[1135954.33, 1297516.1]	$\widehat{b} = 1.14 \times 10^6$	[1134178.33, 1137906.37]
logis	$1,325,950$	-$21,163,576$	$42,327,155$	$42,327,179$	$\widehat{m} = 1.61 \times 10^6$	[1016667.06, 1607560.29]	$\widehat{s} = 1.02 \times 10^6$	[1015054.41, 1017887.2]
norm	$1,325,950$	-$21,750,706$	$43,501,417$	$43,501,441$	$\widehat{\mu} = 2.17 \times 10^6$	[3220115.06, 2161041.21]	$\widehat{\sigma} = 3.22 \times 10^6$	[3216296.98, 3223985.48]



Descriptives for Wages	**Values**
Observations	6,633,449
Sample Mean	1,317,949
Sample Std. Deviation	1,822,342
Minimum	616,000
Median	774,644.6
Maximum	375,726,752
IQR	682,708.3

Figure 7. Diagnostic graphical comparison for the distributions of individual monthly wages (for workers living in municipalities with sizes above $n_{\text{min}} = 287$), fitted by a truncated-lognormal, a Pareto, and a normal distributions, along with some descriptive statistics. Distributions that fit well the data should line up with the black solid line in the Q-Q and P-P plots. Clearly, the normal distribution (green line) is not a good fit for the distribution of monthly wages across workers. Ultimately, the relative best fit among many alternative distributions is given by the smallest AIC, according to which the (truncated) log-normal distribution is the preferred model for monthly wages among Colombian formal workers.

Sizes Figure 8 plots the full empirical complementary cumulative distribution function of municipality sizes. We have followed the methodology proposed by Clauset et al. (2009) to visualize and fit Pareto distributions. We observe in this empirical distribution a natural small-size scale, determined by the estimated minimum size, $\widehat{n_{\text{min}}} \approx 287$ (vertical dashed line), above which the Pareto distribution is well fit (see Clauset et al. 2009 for how to estimate this parameter). We will carry out all our subsequent analyses on the municipalities above $\widehat{n_{\text{min}}}$. Dropping the small-sized municipalities allows us to satisfy the assumption we used for Equation (12), that city sizes are Pareto distributed.[‡] Dropping municipalities that have less than 287 formal workers, means dropping from our analysis $80,526$ workers (only 1.2 percent of total workers in our sample) and 564 municipalities (approximately half of all municipalities).

[‡] Dropping the municipalities with the smallest sizes is typically done as this reduces the potential bias introduced by the fact that their formal employment is overrepresented by public servants whose wages are less determined by economic forces.

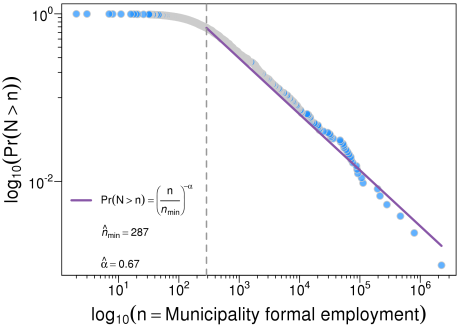
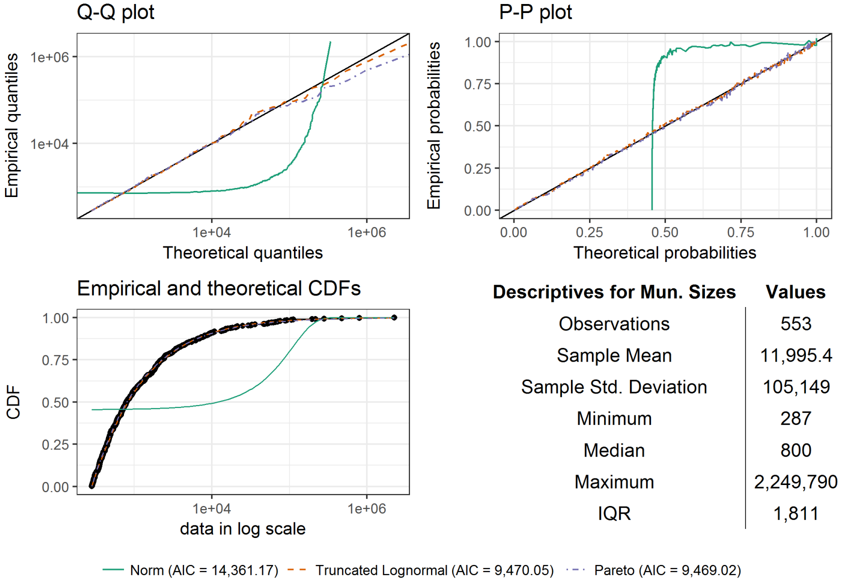


Figure 8. The complementary cumulative empirical distribution of number of workers across municipalities (blue circles) is well-fit by a Pareto distribution (solid purple line).

Table 2. Distributions fitted to municipality sizes. The list of the distributions are ordered from top to bottom by increasing AIC values.

dist	numobs	loglik	AIC	BIC	Parameter 1	C.I.	Parameter 2	C.I.
powerlaw	553	-4, 733.51	9, 469.02	9, 473.33	$\widehat{\alpha} = 0.67$	[0.61, 0.72]		
trunclnorm	553	-4, 733.02	9, 470.05	9, 478.68	$\widehat{\ln(x_0)} = -22.96$	[-50, 0.44]	$\widehat{\sigma} = 6.87$	[3.46, 9.63]
truncweibull	553	-4, 733.08	9, 470.16	9, 478.79	$\widehat{a} = 0.04$	[0.03, 0.11]	$\widehat{b} = 0$	[0, 0]
trunccauchy	553	-4, 755.53	9, 515.06	9, 523.69	$\widehat{l} = 0.001$	[0, 186.88]	$\widehat{s} = 428.74$	[354.78, 531.55]
truncgamma	553	-4, 931.69	9, 867.38	9, 876.02	$\widehat{a} = 0$	[0, 0]	$\widehat{\lambda} = 0.0000$	[0, 0]
lnorm	553	-4, 942.60	9, 889.21	9, 897.84	$\widehat{\ln(x_0)} = 7.16$	[7.04, 7.27]	$\widehat{\sigma} = 1.44$	[1.35, 1.52]
weibull	553	-5, 115.31	10, 234.61	10, 243.24	$\widehat{a} = 0.50$	[0.47, 0.55]	$\widehat{b} = 2,882.57$	[2377.31, 3424.19]
gamma	553	-5, 299.62	10, 603.24	10, 611.87	$\widehat{a} = 0.31$	[0.28, 0.34]	$\widehat{\lambda} = 0.0000$	[0, 0]
truncgumbel	553	-5, 937.73	11, 879.45	11, 888.09	$\widehat{a} = 0.0002$	[0, 3980.75]	$\widehat{b} = 10,684.55$	[9403.29, 11307.02]
trunclogis	553	-6, 003.80	12, 011.61	12, 020.24	$\widehat{m} = 0.0001$	[0, 5064.15]	$\widehat{s} = 10,458.00$	[9169.54, 11040.73]
gumbel	553	-6, 188.90	12, 381.81	12, 390.44	$\widehat{a} = 2,211.40$	[1374.19, 3351.65]	$\widehat{b} = 10,582.62$	[9954.13, 11198.11]
norm	553	-7, 178.59	14, 361.17	14, 369.80	$\widehat{\mu} = 11,995.39$	[4325.07, 22479.46]	$\widehat{\sigma} = 105,053.80$	[100241.97, 111087.68]



Descriptives for Mun. Sizes	Values
Observations	553
Sample Mean	11,995.4
Sample Std. Deviation	105,149
Minimum	287
Median	800
Maximum	2,249,790
IQR	1,811

Figure 9. Diagnostic graphical comparison for the distributions of municipality sizes (with sizes above $n_{\min} = 287$), fitted by a truncated-lognormal, a Pareto, and a normal distributions, along with some descriptive statistics. Distributions that fit well the data should line up with the black solid line in the Q-Q and P-P plots. Clearly, the normal distribution (green line) is not a good fit for the distribution of municipality sizes. Ultimately, the relative best fit among many alternative distributions is given by the smallest AIC, according to which the Pareto distribution is the preferred model for Colombian municipality sizes.