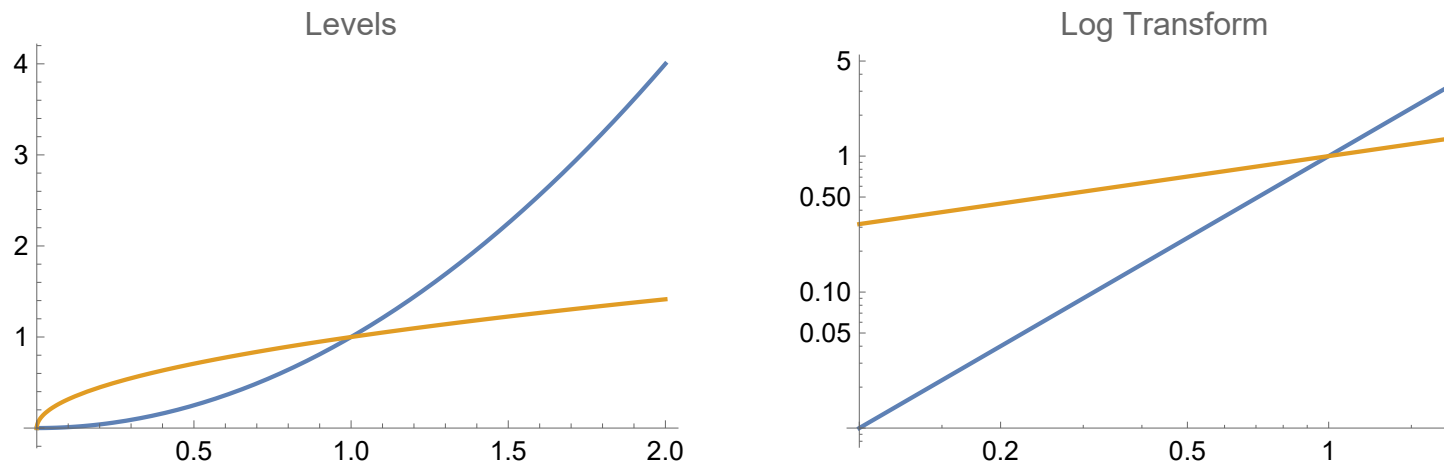
The Pareto Distribution

Background

Power Function

Consider an arbitrary power function, $x \mapsto k\, x^{\alpha}$, where k is a constant and the exponent α governs the relationship. The logarithmic transformation of this power function is linear in Log[x]. That is, if $y = k\, x^{\alpha}$, then $\text{Log}[y] = \text{Log}[k] + \alpha\, \text{Log}[x]$. Another way to say this is that the elasticity of y with respect to x is constant: $\frac{d\, \text{Log}[y]}{d\, \text{Log}[x]} = \alpha$. To communicate the property that the elasticity does not depend on the size of x, the power relationship is called *scale invariant*.

Out[ ]=



Two Power Functions (square vs square root)

Power Law

A power law is a theoretical or empirical relationship governed by a power function. Examples include the following.

- In geometry, the area of a regular polygon is proportional to the square of the length of a side.
- In physics, the gravitational attraction of two objects is inversely proportional to the square of their distance.
- In ecology, Taylor's Law states that the variance of

population density is a power-function of mean population density.

- In economics, Gabaix (1999) finds the population of cities follows a power law (with an inequality parameter close to 1; see below).
- In economics, Luttmer (2007) finds the distribution of employment in US firms follows a power law (with inequality parameter close to 1).
- In economics and business, the Pareto Principle (or 80-20 rule) says that 80% of income accrues to the top 20% of income recipients.

Pareto's Principle

In 1897, Vilfredo Pareto (1848-1923) proposed that the number of people (N_x) with incomes higher than x can be modeled as a power law:

$$N_x = A / x^{\alpha} = A\, x^{-\alpha}$$

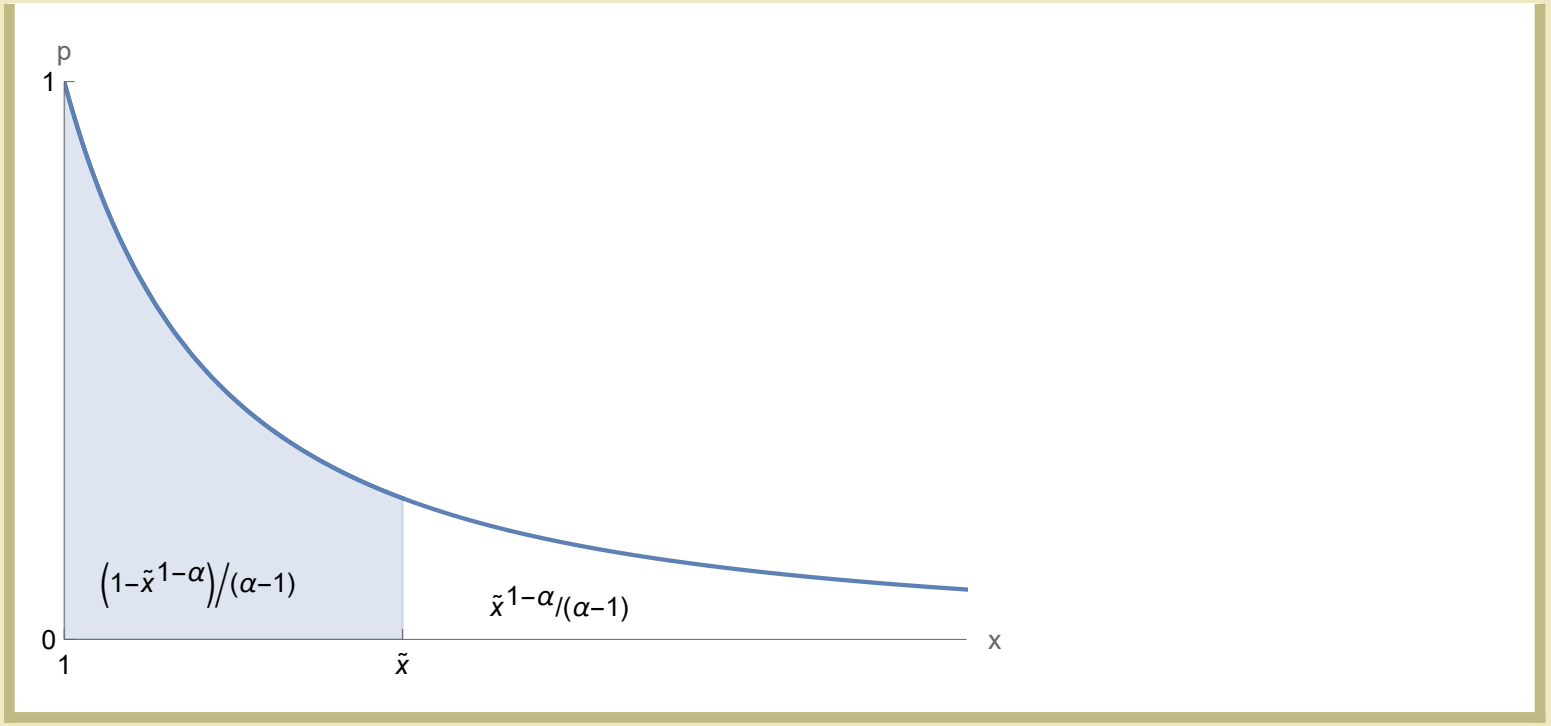
Let the total population be N_0, and let the minimum income be x_0. Then $N_0 = A\ x_0^{-\alpha}$, and we can write this in proportionate terms:

$$N_x / N_0 = (x / x_0)^{-\alpha}$$

Normalize $x_0 = 1$. For ease in discussion, we will

then call x relative income. Using this notation, Pareto's principle is that $n = x^{-\alpha}$ for $x \geq 1$, where $n = N_x / N_0$ is the proportion of relatively rich in the population (i.e., those with relative income greater than x). Assuming $\alpha > 1$, we can find (by integration) that the area under this curve is $1/(\alpha - 1)$, that the proportion of that area that lies before any point x_0 is $1 - x_0^{1-\alpha}$, and that the proportion of that area that lies past any point x_0 is correspondingly $x_0^{1-\alpha}$.

Pareto's Principle: Graphical Illustration



Digression on Integration Details

From the basic principles of integration, we know that the antiderivative of $x^{-\alpha}$ is $x^{1-\alpha} / (1-\alpha)$.

```
Integrate[x^-α, x]
```

$$\frac{x^{1-\alpha}}{1-\alpha}$$

Therefore the definite integral over the interval $[1 \,..\, x_0]$ is $\left(1 - x_0^{1-\alpha}\right) / (1-\alpha)$.

```
Simplify[
  (1 - x0^(1-α))/(α - 1) == Integrate[x^-α, {x, 1, x0}],
  Assumptions → α > 1 && x0 > 1
]
```

```
True
```

Correspondingly, the total area under the curve is $1/(\alpha - 1)$.

```
Integrate[x^-α, {x, 1, ∞}] ~
 Simplify ~ (Assumptions → α > 1)
```

$$\frac{1}{-1+\alpha}$$

The 80-20 Rule

Suppose we are interested in the fraction of total income received by the top 20% of income recipients.

Under the Pareto principle that $n = x^{-\alpha}$, we have seen that the share of total income received by those with relative incomes above x can then be written as $s = x^{1-\alpha}$.

We can also invert the Pareto principle to yield $x = n^{-1/\alpha}$.

Combining these two observations, for a given

proportion n of top income earners, we can find the associated top-share of income as $s = \left(n^{-1/\alpha}\right)^{1-\alpha} = n^{1-1/\alpha}$.

```
Eliminate[n == x^-α && s == x^(1-α), x] // Quiet
Solve[%, α] // PowerExpand // Simplify
```

$$s^{\frac{1}{1-\alpha}} == \left(\frac{1}{n}\right)^{\frac{1}{\alpha}}$$

$$\left\{\left\{\alpha \to \frac{\text{Log}[n]}{\text{Log}[n] - \text{Log}[s]}\right\}\right\}$$

For example, for Pareto's 80-20 rule to hold, we must have $\alpha \approx 1.16$.

$$\frac{\text{Log}[n]}{\text{Log}[n] - \text{Log}[s]} \text{ /. } \{s \to 0.80, n \to 0.20\}$$

```
1.16096
```

Deducing the Income Share of the Richest

Given α we can compute the total income share of the proportion p of the richest income recipients.

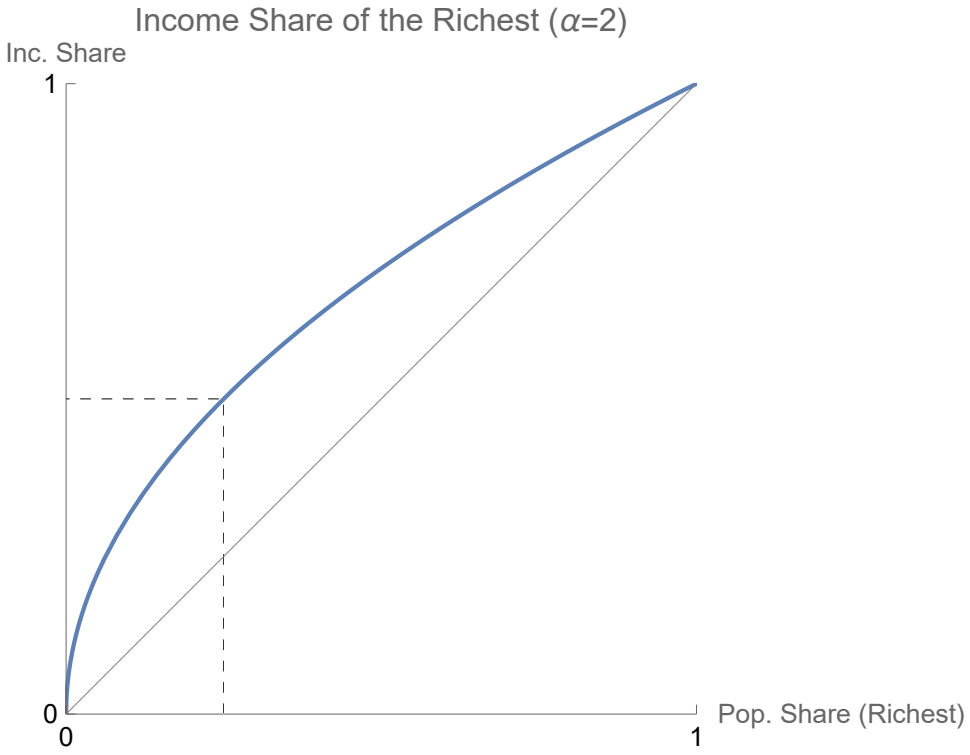
For example, if α == 2 then the richest 1% of income recipients receive 10% of total income.

```
p^(1-1/α) /. {p → 0.01, α → 2}
```

```
0.1
```

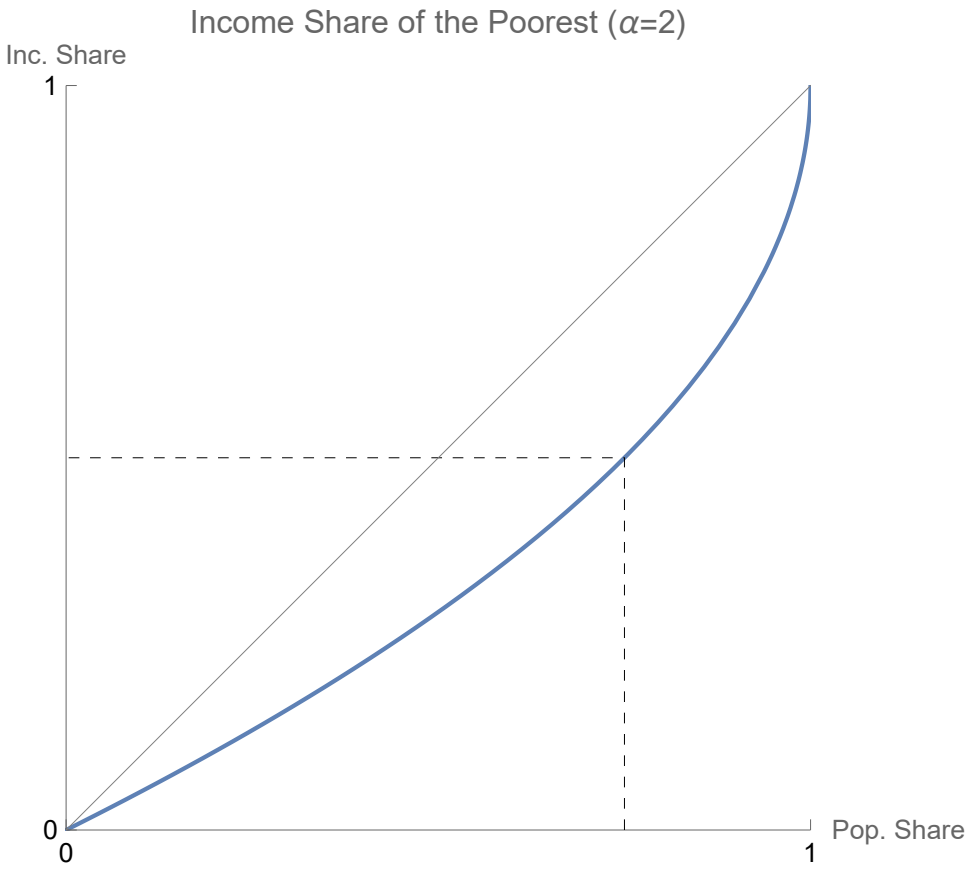
It naturally follows that the rest of the population (i.e., the poorest 99% of income recipients) receive the rest of income (i.e., 90% of total income).

Income Share of the Richest: Illustrated



Income Share of the Poorest (Lorenz Curve)

It is somewhat more common to rotate this plot 180 ° to display the income share of the poorest. The result is usually called a Lorenz curve. (Note however Lorenz (1905) plotted the cumulative population share against the cumulative wealth share. I.e., our Lorenz curve is the inverse of his—a reflection through the 45 ° line.)


Income Share of the Poorest (α=2)
Inc. Share
1
0
0
1
Pop. Share

Pareto Inequality

It is common to report $\eta = 1/\alpha$ as the measure of Pareto inequality. Since we generally think $\alpha \in (1, 2)$, we should find $\eta \in (0.5, 1)$. Jones (2015) claims that in the contemporary US, $\eta \approx 0.6$ for the distribution of income. This should correspond to a top 1% share of around 16%.

```
p^(1-1/α) /. {p → 0.01, α → 10/6}
```

```
0.158489
```

Power-Law Distributions

Continuous Power-Law Probability Distribution

Define a continuous power-law distribution with shape parameter $\alpha>0$ and size parameter $x_0 > 0$ to be a distribution with probability distribution function $p[x] = k\, x^{-(1+\alpha)}$ for $x > x_0$. The antiderivative is

$$-\frac{k\, x^{-\alpha}}{\alpha}$$

We can therefore compute the (improper) integral

over $[x_0 \,..\, \infty]$ to be

$$\frac{k\, x0^{-\alpha}}{\alpha}$$

The constant k must be chosen to satisfy normalization (unitarity), since the total area under the PDF must equal 1. This means that k must be

$$x0^{\alpha}\, \alpha$$

Plugging in our solution for the constant of integration back into our PDF, we fully characterize of our power-law distribution in terms of two parameters: the shape parameter (α) and the size parameter (x_0).

$x^{-1-\alpha}\, x0^{\alpha}\, \alpha$

This distribution is usually known as the Pareto distribution, and we will soon relate it to the Pareto principle. (It is sometimes known as the Bradford distribution, after Bradford (1934), but this term also refers to a related truncated distribution.)

The Pareto Distribution

The social sciences have found that the Pareto distribution embodies a useful power law. The Pareto Distribution is most often presented in terms of its survival function, which gives the probability of seeing larger values than x. (This is often known as the complementary CDF, since it is 1-CDF. It is sometimes called the reliability function or the tail function.) The survival function of a Pareto distribution for $x \in [x_0 \,..\, \infty]$ is

$$\left(\frac{x}{x0}\right)^{-\alpha}$$

This value of this survival function is initially 1 and declines to 0 as x increases. It defines a continuous probability distribution on $[x_0 \,..\, \infty]$.

We are only interested in $x > x_0$, and we are usually interested in $\alpha > 1$ (which is required for finite mean value). We call $x_0 > 0$ the location parameter; we call $\alpha > 0$ the shape parameter (or slope parameter, or Pareto index); and we say the distribution is Pareto$[x_0, \alpha]$.

Survival: Intuition

Consider a population of households and suppose sampling household incomes is like sampling from a Pareto[10000,2].
What proportion of people earn more than $100000 (i.e., ten times the minimum)? From the form of the survival function $(x/x_0)^{-2}$, it should be obvious that the answer is 10^{-2}: only 1 in 100 households earn more than $100000.

Elaborating, we see that for $\alpha = 2$ and any x_0, we find

that 1% of the population has income greater than $10 * x_0$. This is one way in which the Pareto distribution (along with other power law distributions) is *scale free*.

```
Simplify[
 SurvivalFunction[
  ParetoDistribution[x0, 2], 10 * x0],
 Assumptions → x0 > 0]
```

$$\frac{1}{100}$$

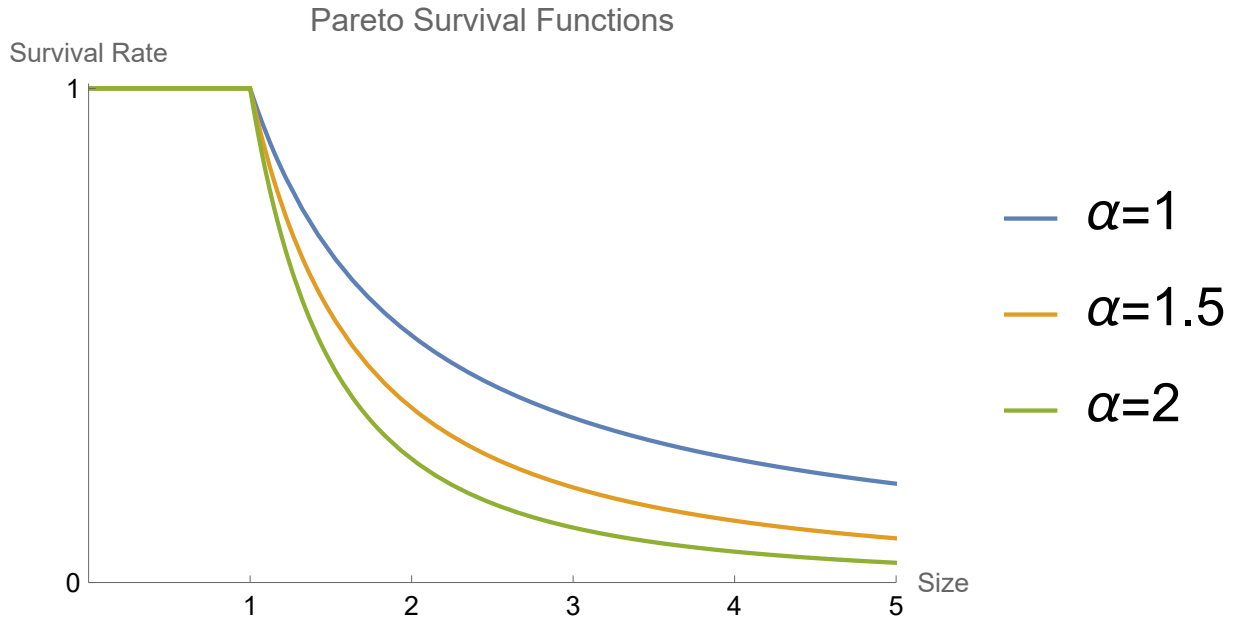
What is more, this scale-free relationship holds as well for subgroups: only 1% of the top 1% will have incomes that are another ten times higher.

```
Simplify[
 SurvivalFunction[
  ParetoDistribution[x0, 2], 10 * 10 * x0],
 Assumptions → x0 > 0]
```

$$\frac{1}{10\,000}$$

Survival and the Pareto Index

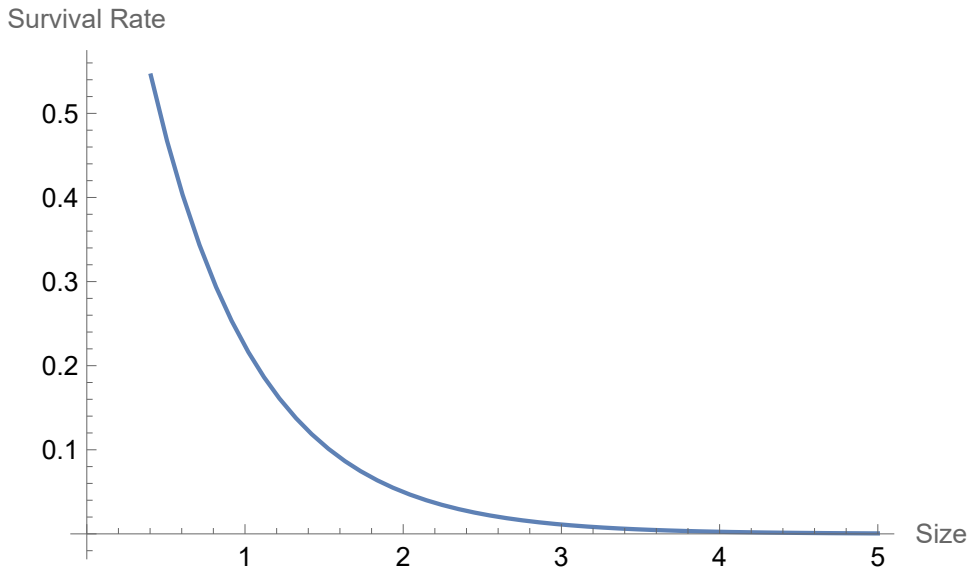
To simplify comparisons, let us work with a normalized Pareto distribution: $S[x] = x^{-\alpha}$ for $x \geq 1$. That is, we normalize $x_0 = 1$. This lets us work with a single parameter: α, the shape parameter of the Pareto distribution.


Pareto Survival Functions
Survival Rate
1
0
1
2
3
4
5
Size
α=1
α=1.5
α=2

Contrast with the Exponential Distribution

The Exponential Distribution

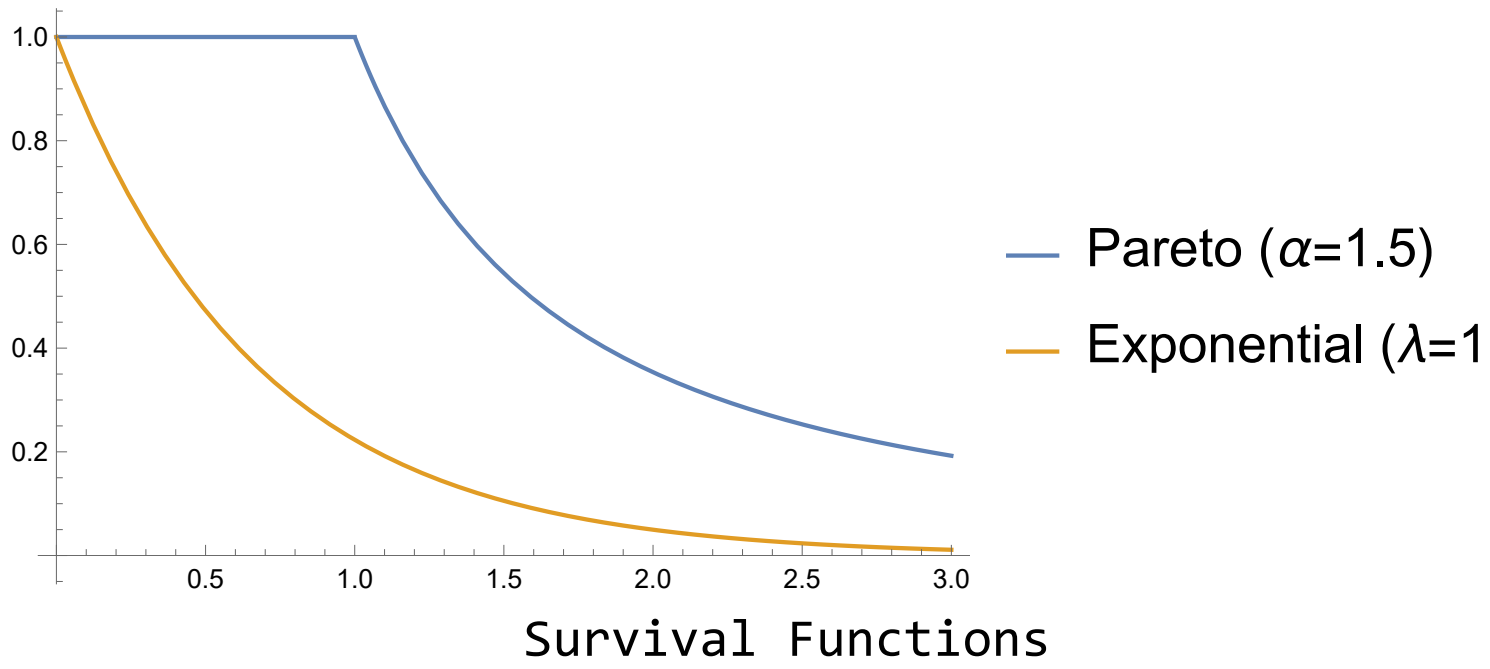
Let us briefly compare the Pareto distribution to the exponential distribution, which may initially seem similar. The survival function of the exponential distribution is $S[x] = e^{-\lambda x}$ for $x \geq 0$, where $\lambda > 0$ is the shape parameter of the distribution.
(Correspondingly the CDF is $F[x] = 1 - e^{-\lambda x}$ and the PDF is $f[x] = \lambda e^{-\lambda x}$.)



Survival Function (Exponential[1.5])

Contrast with the Exponential Distribution

Because the survival rate in the tail is higher for the Pareto distribution than for the exponential, we say that the Pareto distribution has a fat tail. We can begin to see the difference by plotting the survival functions.


1.0
0.8
0.6
0.4
0.2
0.5
1.0
1.5
2.0
2.5
3.0
Pareto (α=1.5)
Exponential (λ=1
Survival Functions

Further Contrast with the Exponential Distribution

Recall that the survival function of the exponential distribution is $S[x] = e^{-\lambda x}$. At first sight the Pareto distribution may seem to have much in common with the exponential distribution. However, the survival rate of the Pareto distribution declines much more slowly. Here is a way to consider that contrast: for x1, x2 > x_0 and associated N1, N2, the Pareto distribution implies

$$\log(N_1/N_2) = -\alpha \log(x_1/x_2)$$

whereas for the exponential distribution

$$\log(N_1/N_2) = -\lambda(x_1 - x_2)$$

Under a Pareto distribution, relative survival depends only on the ratio (x_1/x_2), so the same relationship holds anywhere tail of the income distribution, no matter how far out. If the top 20% of people receive 80% of income, then the top 4% (20% of 20%) receive 64% (80% of 80%) of income, and so on.

Fat-Tailed Distributions

Most popular probability distributions have well defined means, variances, and higher-order moments. For example, the exponential distribution with parameter $\lambda > 0$ has a mean of $1/\lambda$ and a variance of $1/\lambda^2$. For such distributions, outcomes far from the mean are very rare. Other distributions have "fat" tails: outcomes far from the mean are less rare. For example, the Pareto distribution has infinite variance if $\alpha \leq 2$.

A probability distribution is said to be *fat-tailed* if

eventually (i.e., as x gets big) the PDF is proportional to a power function of the form $x \mapsto x^{-(1+\alpha)}$ where $\alpha > 0$.

Equivalently, the survival function $P[X > x]$ is eventually proportional to a power function of the form $x \mapsto x^{-\alpha}$ where $\alpha > 0$. Contrast this with the survival function for an exponential distribution: $e^{-\lambda x}$. No matter how small we make $\lambda > 0$, we will find $e^{-\lambda x} \Big/ x^{-\alpha}$ is eventually tiny. Any power law distribution eventually has a much bigger tail than any exponential distribution.

```
Assuming[λ > 0 && α > 0,
 Limit[Exp[-λ x] / x^-α, x → ∞]
]
```

```
0
```

Algebraic Details (Fat Tail)

We can show more formally that survival function declines more rapidly for the exponential than for the Pareto distribution.

$\lim_{x\to\infty} e^{-\lambda x+\alpha \ln x} = \text{Exp}[\lim_{x\to\infty} -\lambda x + \alpha \ln x]$ by continuity, and we can write the latter as $\text{Exp}[\lim_{x\to\infty} (-\lambda x/\ln x + \alpha) \ln x]$. Recall that $\infty == \lim_{x\to\infty} (x/\ln x)$, or apply L'Hospital's rule to show it. So we have

```
Limit[(-λ x + α) Log[x], x → ∞] == -∞
```

and correspondingly

```
Exp[Limit[(-λ x + α) Log[x], x → ∞]] == 0
```

Intimate Relation of Pareto to Exponential

There is an intimate relationship between the Pareto and exponential distributions. Recall that the survival function of the exponential distribution is $e^{-\lambda x}$.

Let Y be exponentially distributed, with survival function $P[Y > y] = (e^{y})^{-\alpha}$. (For convenience we have changed the name of the shape parameter.) Define a new random variable by $X = x_0 e^{Y}$. Then compute the survival probability

P[X>x]=P[x0

$$e^{Y} > x\Big] = P\Big[e^{Y} > x/x0\Big] = P[Y > \text{Log}[x/x0]] = (x/x0)^{-\alpha}$$

Comparing to the resulting survival function for the Pareto distribution, we see that X has a Pareto distribution.

```
Simplify[
 SurvivalFunction[ExponentialDistribution[α]][
  Log[x / x0]],
 x ≥ x0 > 0
]
```

$$\left(\frac{x0}{x}\right)^{\alpha}$$

Some Pareto Details

Pareto Distribution: CDF

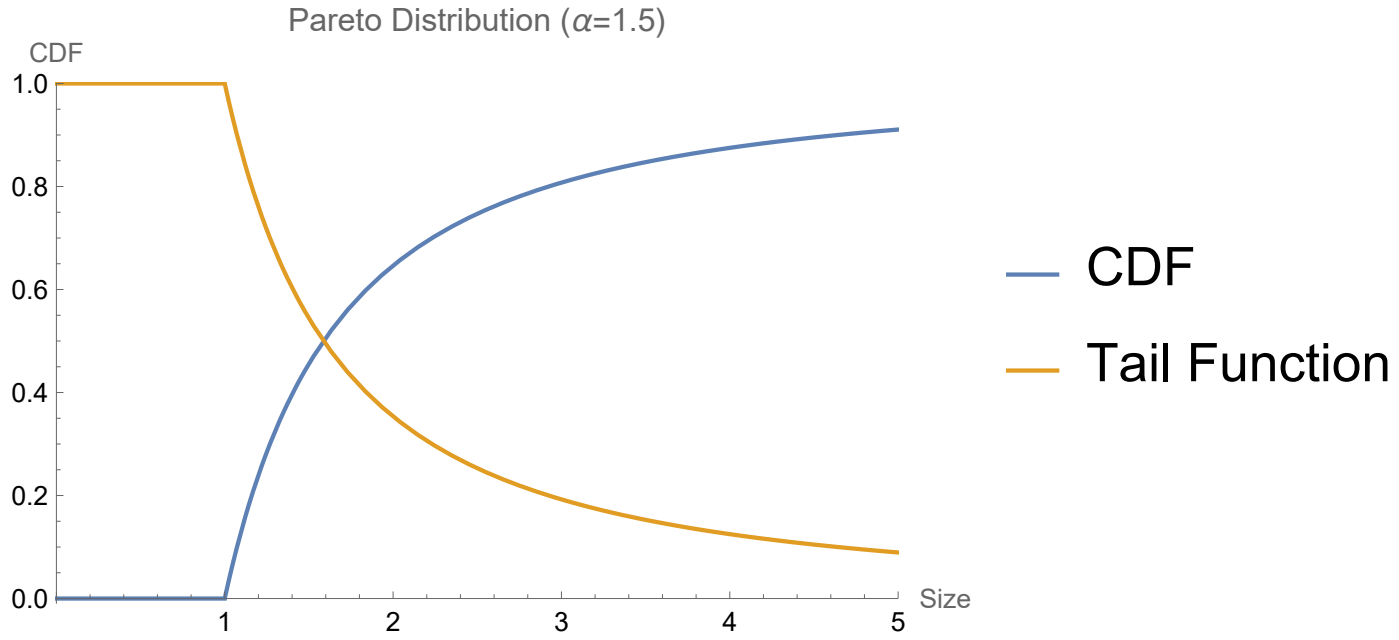
Recall the that the Pareto survival function is $(x/x_0)^{-\alpha}$ for $x \geq x_0$. (The minimum value x_0 is called the location parameter; α is called the shape parameter.) The survival function returns the probability of seeing an outcome bigger than x. Since the cumulative distribution function (CDF) gives the probability of seeing a values less than or equal to x, it must be $1 - (x/x_0)^{-\alpha}$.

```
Simplify[
 CDF[ParetoDistribution[x0, α], x],
 Assumptions → x ≥ x0 > 0]
```

$$1 - \left(\frac{x0}{x}\right)^{\alpha}$$

CDF vs Survival Function of Pareto Distributions

Recall the that the Pareto survival function for values greater than the minimum is $(x/x_0)^{-\alpha}$. Normalize $x_0 == 1$ for presentational simplicity, so that we have $n == x^{-\alpha}$.


Pareto Distribution (α=1.5)
CDF
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
Size
CDF
Tail Function

PDF of Pareto Distributions

The probability distribution function (PDF) gives the likelihood of an outcome occurring near any particular possible value.

```
Simplify[
 PDF[ParetoDistribution[x0, α], x],
 Assumptions → x > x0 > 0]
```

$$x^{-1-\alpha}\, x0^{\alpha}\, \alpha$$

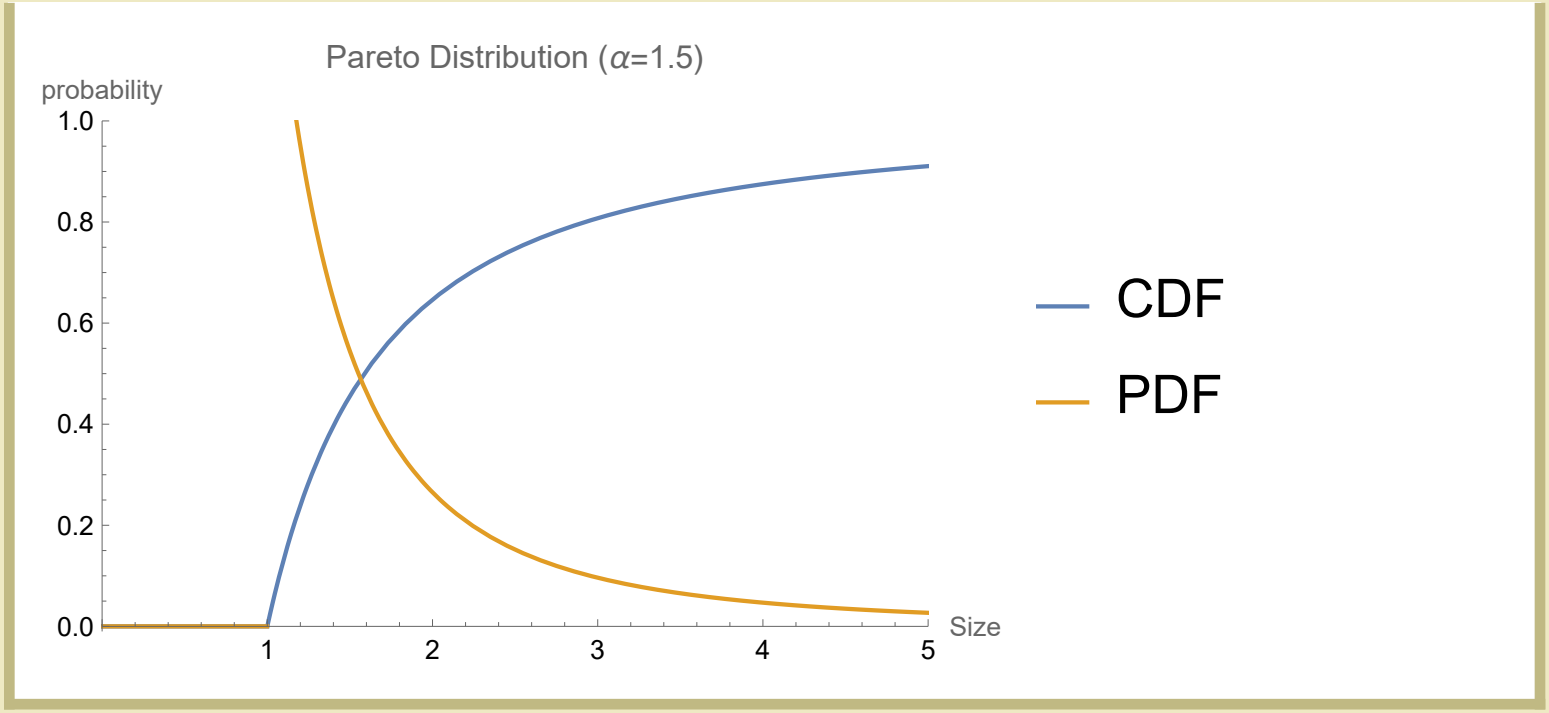
We may say the the PDF gives us the marginal rate

at which the cumulative probability (given by the CDF) is increasing. That is, the PDF is the derivative of the CDF.

```
Simplify[
 PDF[ParetoDistribution[x0, α], x] ==
  D[CDF[ParetoDistribution[x0, α], x], x],
 Assumptions → x > x0 > 0
]
```

```
True
```

CDF and PDF of Pareto Distributions: Illustration



Log-Linear Survival

Recall that the survival function of the normalized Pareto can be characterized by $n == x^{-\alpha}$. Take the logarithm of both sides:

```
Log /@ (n == x^-α) // PowerExpand
 (* 11.3+, use ApplySides instead of Map *)
```

$$\text{Log}[n] == -\alpha\, \text{Log}[x]$$

Note that we have a linear relationship in logs. We can therefore say that the size elasticity of the

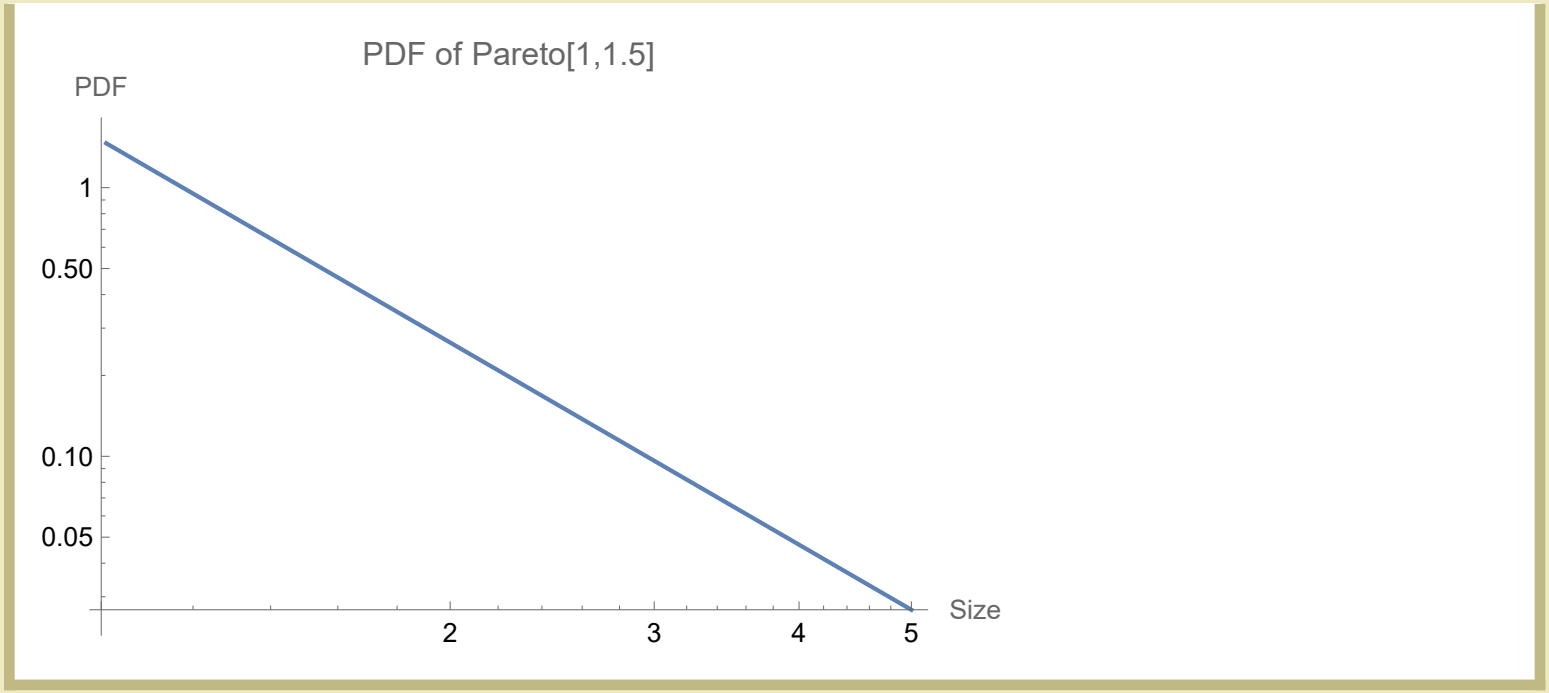
survival rate is the constant α. (We will return to this.)

Similarly, since the CDF is $1 - x^{-\alpha}$, the PDF is $\alpha\, x^{-(\alpha+1)}$, which a log transformation again linearizes.

```
Simplify[
  Log[α x^-(α+1)],
  Assumptions → α > 0 && x > 0] // PowerExpand
```

```
(-1 - α) Log[x] + Log[α]
```

PDF of Pareto Distributions: Loglinearity



Average Size

The PDF of the Pareto[x0,α] distribution is $(\alpha / x_0)\,(x / x_0)^{-(\alpha+1)}$.

Use this PDF to compute the average size of a draw from the Pareto distribution as $\int_{x0}^{\infty} x\, p(x)\, dx == \int_{x0}^{\infty} \alpha \left(\frac{x}{x_0}\right)^{-\alpha} dx$. This integral only exists for $\alpha > 1$, in which case it is

$$\frac{x0\,\alpha}{-1+\alpha}$$

We can also compute this result using WL's builtin functions (although this provides less conceptual insight).

```
Simplify[
 Mean@ParetoDistribution[x0, α],
 x0 > 0 && α > 1]
```

$$\frac{x0\,\alpha}{-1+\alpha}$$

Proportion of Total Income

Suppose income follows a Pareto[x0,α] distribution where we let $p[x]=\alpha\, x0^{\alpha}\, x^{-(1+\alpha)}$ is the PDF. Then the weighted sum of all incomes less than or equal to any given cutoff $t \geq x_0$ is

$$\int_{x0}^{t} x\, p[x]\, dx == \frac{x0\,\alpha}{-1+\alpha} - \frac{t^{1-\alpha}\, x0^{\alpha}\,\alpha}{-1+\alpha}$$

Dividing by the mean (i.e., the probability weighted sum of all incomes, which we computed above) produces an expression for the proportion of total

income constituted by incomes of t or less: $1 - (t/x0)^{1-\alpha}$.

```
(x0 α/(-1 + α) - t^(1-α) x0^α α/(-1 + α)) / (x0 α/(-1 + α)) // Simplify
```

```
1 - t^(1-α) x0^(-1+α)
```

Note that this expression only makes sense for $\alpha > 1$ (the case in which the mean exists). Also note that the value of the result $1 - (t/x0)^{1-\alpha}$ depends only on the value of the ratio t/x_0.

Lorenz Curve and Gini Coefficient

Lorenz Curve: Parametric Plot

A Lorenz curve for x (e.g., income or wealth) plots the cumulative share of x vs the cumulative share of the population possessing it.

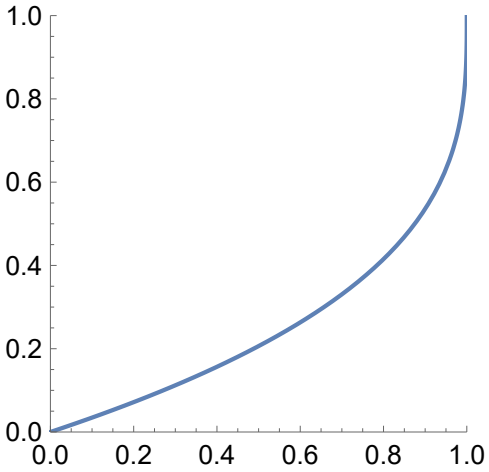
We have just found that the cumulative share of income for incomes up to t can be written as $1 - (t/\text{x0})^{1-\alpha}$.

Additionally, we have see that the CDF at income t (which gives the proportion of the population earning no more than t) is $1 - (t/\text{x0})^{-\alpha}$.

So for any alpha, we can make a parametric plot of

the Lorenz curve. Defining $m = x0/t$ we can write:

```
With[{α = 1.5},
 ParametricPlot[{1 - m^α, 1 - m^(α-1)}, {m, 0, 1},
  PlotRange → {{0, 1}, {0, 1}},
  AspectRatio → 1, ImageSize → Small]]
```



```
ClearAll[nexti]
Module[{prev, this}, prev = 0;
 nexti[] := (this = prev + 1;
   prev = this)]
nexti[]
nexti[]
```

```
1
```

```
2
```

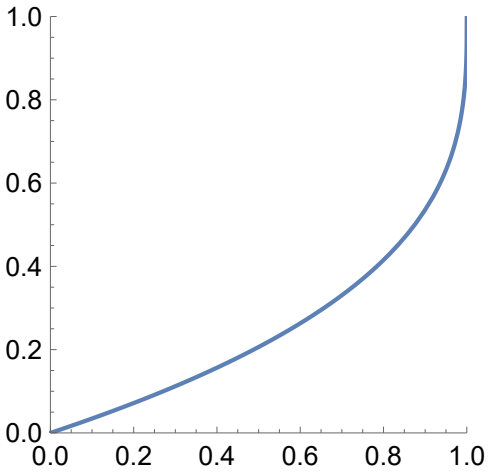
```
nexti[]
```

```
3
```

Another Approach

Of course, from the parametric formulation we can solve for the income share as a function of the population share.

```
With[{α = 1.5},
 Plot[1 - (1 - p)^(1 - 1/α), {p, 0, 1},
  PlotRange → {{0, 1}, {0, 1}},
  AspectRatio → 1, ImageSize → Small]]
```



Lorenz Curve

We can transform the above parametric representation of the Lorenz curve to express the cumulative share of income as a function of the cumulative share of the population earning it. Recall that the CDF at income t gives the share of the population (say, $s(t)$) earning less than t. Since this CDF is strictly increasing, we can produce the inverse function $t(s)$,

```
cdfPareto /. {x → t}
ts =
 Solve[s == %, t] // Flatten (* inverse of CDF *)
```

$$1 - \left(\frac{x0}{t}\right)^{\alpha}$$

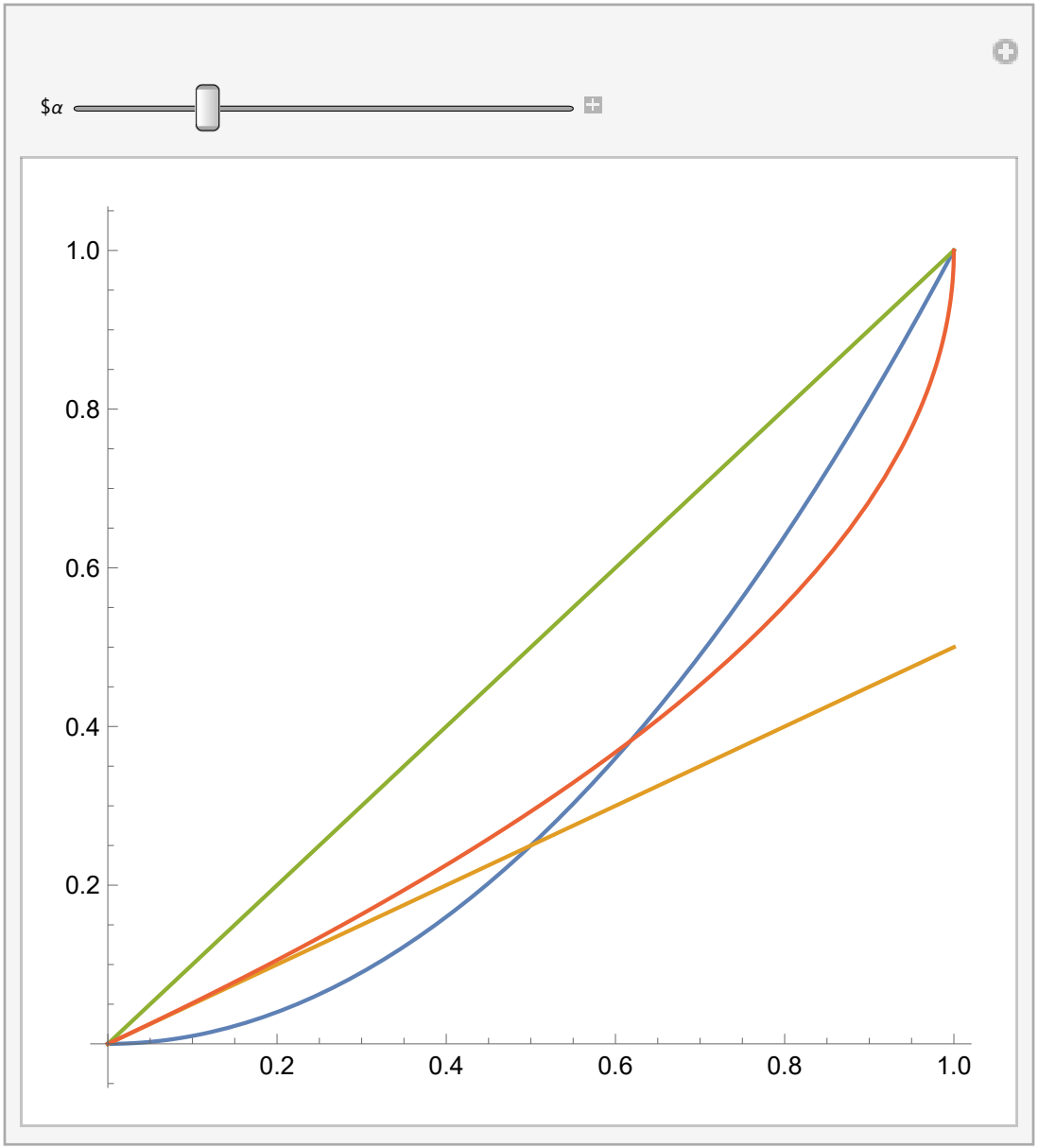
Solve::ifun : Inverse functions are being used by Solve, so some solutions may not be found; use Reduce for complete solution information. >>

$$\left\{t \to (1 - s)^{-1/\alpha}\, x0\right\}$$

which we can then substitute into our expression of the cumulative share of income as a function of the income level to get the cumulative share of income as a function of the cumulative share of the population.

```
lorenzPareto = Simplify[cumShare /. ts,
  Assumptions → x0 > 0 && α > 1]
Manipulate[
 Plot[{s^$α, ($α - 1)/$α s, s, 1 - (1 - s)^(($α - 1)/$α)},
  {s, 0, 1}, AspectRatio → 1],
 {{$α, 2}, 1.01, 5}]
```

```
cumShare
```


$α
1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0

Approximation

Let s_y be the cumulative share of income and s_p be the cumulative share of the population. We have found

$$s_y == 1 - (1 - s_p)^{\delta}$$

$$\log(1 - s_y) == \delta \log(1 - s_p)$$

$$s_y \approx \delta\, s_p$$

Discrete Approximation of the Lorenz Curve

We can equivalently parameterize the LC by $\delta = 1 - 1/\alpha$:

```
δ = 1 - 1 / α
lorenzPareto == 1 - (1 - s)^δ
Assuming[0 < s < 1,
 Solve[1 - (1 - s)^d1 == s^d2, {d1, d2}]]
```

$$1 - \frac{1}{\alpha}$$

$$\text{cumShare} == 1 - (1 - s)^{1 - \frac{1}{\alpha}}$$

Solve::ifun : Inverse functions are being used by Solve, so some solutions may not be found; use Reduce for complete solution information. >>

Solve::svars : Equations may not give solutions for all "solve" variables. >>

$$\left\{\left\{d2 \to \frac{\text{Log}\left[1 - (1 - s)^{d1}\right]}{\text{Log}[s]}\right\}\right\}$$

Suppose we have N equal-sized bins of income recipients. Let $p_i = 1/N$ be the i-th incremental proportion of the population, so that the income share

earned by the *i*-th bin is

```
Clear[i, n, δ]
ishare[i_] := (1 - (i - 1)/n)^δ - (1 - i/n)^δ
```

So for example the income shares for 100 bins with $\alpha = 2$ will be

```
n = 100
nvals = Range[n]
cumsharePop = nvals / n
ishares =
 Map[ishare, Range[n]] /. {δ → 1 - 1 / α} /. {α → 2.}
ListPlot[Transpose[{nvals / n, ishares}],
 PlotLabel → "rising incremental income"]
```

```
cumshareInc = Accumulate[ishares]
cumshareInc == (lorenzPareto /. {s → cumsharePop} /.
    {δ → 1 - 1 / α} /. {α → 2.})
ListPlot[Transpose[{cumsharePop, cumshareInc}]]
estB = Total[(n + 1 - nvals) * ishares] /
  (n * Total[ishares])
estG = 1 - 2 * estB
actualG = 1 / (2 * α - 1) /. {α → 2.}
```

```
100
```

```
{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,
 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26,
 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38,
 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50,
 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62,
 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75,
 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88,
 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100}
```

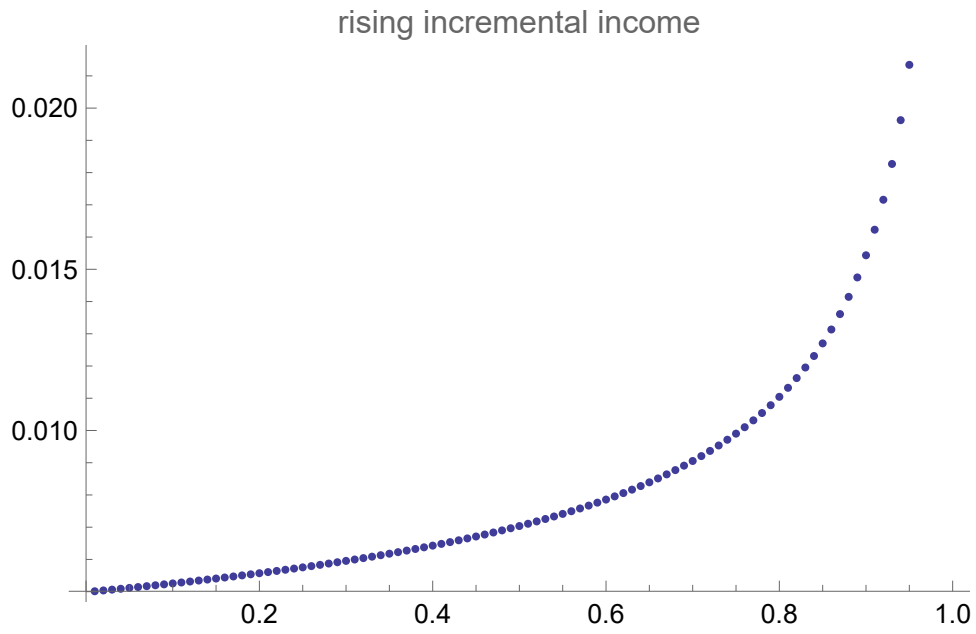
$$\left\{\frac{1}{100}, \frac{1}{50}, \frac{3}{100}, \frac{1}{25}, \frac{1}{20}, \frac{3}{50}, \frac{7}{100}, \frac{2}{25}, \frac{9}{100},\right.$$
$$\frac{1}{10}, \frac{11}{100}, \frac{3}{25}, \frac{13}{100}, \frac{7}{50}, \frac{3}{20}, \frac{4}{25}, \frac{17}{100}, \frac{9}{50}, \frac{19}{100},$$
$$\left.\frac{1}{5}, \frac{21}{100}, \frac{11}{50}, \frac{23}{100}, \frac{6}{25}, \frac{1}{4}, \frac{13}{50}, \frac{27}{100}, \frac{7}{25}, \frac{29}{100},\right.$$

$$\left\{\frac{3}{10}, \frac{31}{100}, \frac{8}{25}, \frac{33}{100}, \frac{17}{50}, \frac{7}{20}, \frac{9}{25}, \frac{37}{100}, \frac{19}{50}, \frac{39}{100},\right.$$
$$\frac{2}{5}, \frac{41}{100}, \frac{21}{50}, \frac{43}{100}, \frac{11}{25}, \frac{9}{20}, \frac{23}{50}, \frac{47}{100}, \frac{12}{25}, \frac{49}{100},$$
$$\frac{1}{2}, \frac{51}{100}, \frac{13}{25}, \frac{53}{100}, \frac{27}{50}, \frac{11}{20}, \frac{14}{25}, \frac{57}{100}, \frac{29}{50}, \frac{59}{100},$$
$$\frac{3}{5}, \frac{61}{100}, \frac{31}{50}, \frac{63}{100}, \frac{16}{25}, \frac{13}{20}, \frac{33}{50}, \frac{67}{100}, \frac{17}{25}, \frac{69}{100},$$
$$\frac{7}{10}, \frac{71}{100}, \frac{18}{25}, \frac{73}{100}, \frac{37}{50}, \frac{3}{4}, \frac{19}{25}, \frac{77}{100}, \frac{39}{50}, \frac{79}{100}, \frac{4}{5},$$
$$\frac{81}{100}, \frac{41}{50}, \frac{83}{100}, \frac{21}{25}, \frac{17}{20}, \frac{43}{50}, \frac{87}{100}, \frac{22}{25}, \frac{89}{100}, \frac{9}{10},$$
$$\left.\frac{91}{100}, \frac{23}{25}, \frac{93}{100}, \frac{47}{50}, \frac{19}{20}, \frac{24}{25}, \frac{97}{100}, \frac{49}{50}, \frac{99}{100}, 1\right\}$$

```
{0.00501256, 0.00503794, 0.00506371,
 0.00508988, 0.00511646, 0.00514346, 0.0051709,
```

```
0.00519877, 0.0052271, 0.0052559, 0.00528518,
0.00531496, 0.00534525, 0.00537606, 0.0054074,
0.00543931, 0.00547178, 0.00550484, 0.00553851,
0.00557281, 0.00560775, 0.00564336, 0.00567965,
0.00571665, 0.00575438, 0.00579288, 0.00583215,
0.00587224, 0.00591316, 0.00595495, 0.00599764,
0.00604126, 0.00608585, 0.00613144, 0.00617807,
0.00622577, 0.00627461, 0.00632461, 0.00637582,
0.0064283, 0.00648209, 0.00653726, 0.00659387,
0.00665197, 0.00671163, 0.00677293, 0.00683593,
0.00690073, 0.00696741, 0.00703606, 0.00710678,
0.00717968, 0.00725486, 0.00733246, 0.00741261,
0.00749544, 0.00758111, 0.00766978, 0.00776165,
0.00785689, 0.00795573, 0.0080584, 0.00816515,
0.00827625, 0.00839202, 0.00851279, 0.00863892,
```

```
0.00877084, 0.00890899, 0.00905388, 0.00920608,
0.00936622, 0.00953502, 0.00971329, 0.00990195,
0.0101021, 0.0103148, 0.0105416, 0.010784,
0.011044, 0.0113237, 0.0116258, 0.0119535,
0.0123106, 0.0127017, 0.0131326, 0.0136106,
0.014145, 0.0147477, 0.0154347, 0.0162278,
0.0171573, 0.0182676, 0.0196262, 0.0213422,
0.0236068, 0.0267949, 0.0317837, 0.0414214, 0.1}
```



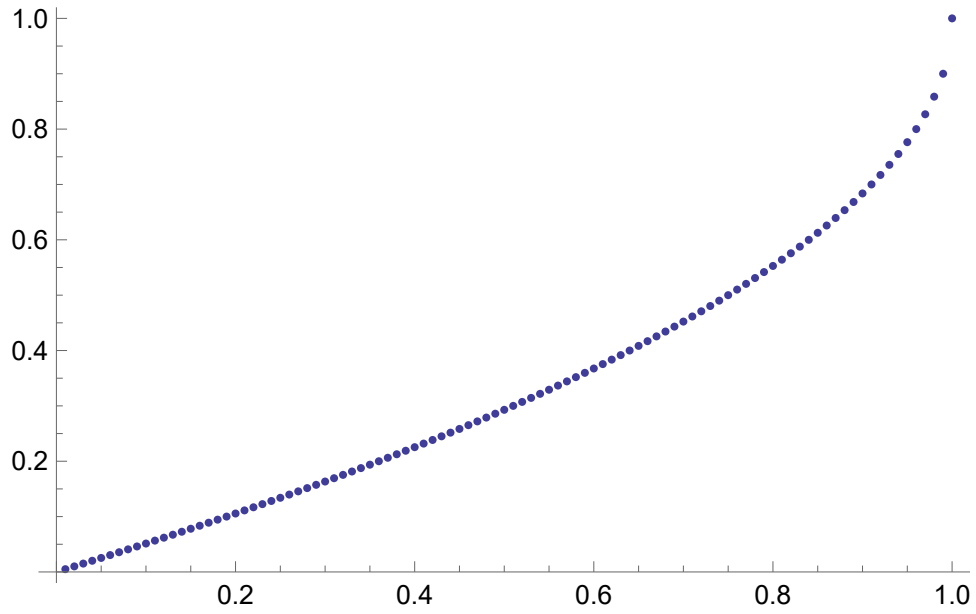

```
{0.00501256, 0.0100505, 0.0151142,
 0.0202041, 0.0253206, 0.030464, 0.0356349,
 0.0408337, 0.0460608, 0.0513167, 0.0566019,
 0.0619168, 0.0672621, 0.0726382, 0.0780456,
 0.0834849, 0.0889566, 0.0944615, 0.1, 0.105573,
 0.111181, 0.116824, 0.122504, 0.12822, 0.133975,
```

```
0.139767, 0.1456, 0.151472, 0.157385, 0.16334,
0.169338, 0.175379, 0.181465, 0.187596,
0.193774, 0.2, 0.206275, 0.212599, 0.218975,
0.225403, 0.231885, 0.238423, 0.245017, 0.251669,
0.25838, 0.265153, 0.271989, 0.27889, 0.285857,
0.292893, 0.3, 0.30718, 0.314435, 0.321767,
0.32918, 0.336675, 0.344256, 0.351926, 0.359688,
0.367544, 0.3755, 0.383559, 0.391724, 0.4,
0.408392, 0.416905, 0.425544, 0.434315, 0.443224,
0.452277, 0.461484, 0.47085, 0.480385, 0.490098,
0.5, 0.510102, 0.520417, 0.530958, 0.541742,
0.552786, 0.56411, 0.575736, 0.587689, 0.6,
0.612702, 0.625834, 0.639445, 0.65359, 0.668338,
0.683772, 0.7, 0.717157, 0.735425, 0.755051,
0.776393, 0.8, 0.826795, 0.858579, 0.9, 1.}
```

```
{0.00501256, 0.0100505, 0.0151142, 0.0202041,
```

```
0.0253206, 0.030464, 0.0356349, 0.0408337,
0.0460608, 0.0513167, 0.0566019, 0.0619168,
0.0672621, 0.0726382, 0.0780456, 0.0834849,
0.0889566, 0.0944615, 0.1, 0.105573, 0.111181,
0.116824, 0.122504, 0.12822, 0.133975, 0.139767,
0.1456, 0.151472, 0.157385, 0.16334, 0.169338,
0.175379, 0.181465, 0.187596, 0.193774, 0.2,
0.206275, 0.212599, 0.218975, 0.225403, 0.231885,
0.238423, 0.245017, 0.251669, 0.25838, 0.265153,
0.271989, 0.27889, 0.285857, 0.292893, 0.3,
0.30718, 0.314435, 0.321767, 0.32918, 0.336675,
0.344256, 0.351926, 0.359688, 0.367544, 0.3755,
0.383559, 0.391724, 0.4, 0.408392, 0.416905,
0.425544, 0.434315, 0.443224, 0.452277,
0.461484, 0.47085, 0.480385, 0.490098, 0.5,
```

```
0.510102, 0.520417, 0.530958, 0.541742, 0.552786,
0.56411, 0.575736, 0.587689, 0.6, 0.612702,
0.625834, 0.639445, 0.65359, 0.668338, 0.683772,
0.7, 0.717157, 0.735425, 0.755051, 0.776393,
0.8, 0.826795, 0.858579, 0.9, 1.} == cumShare
```



```
0.338537
```

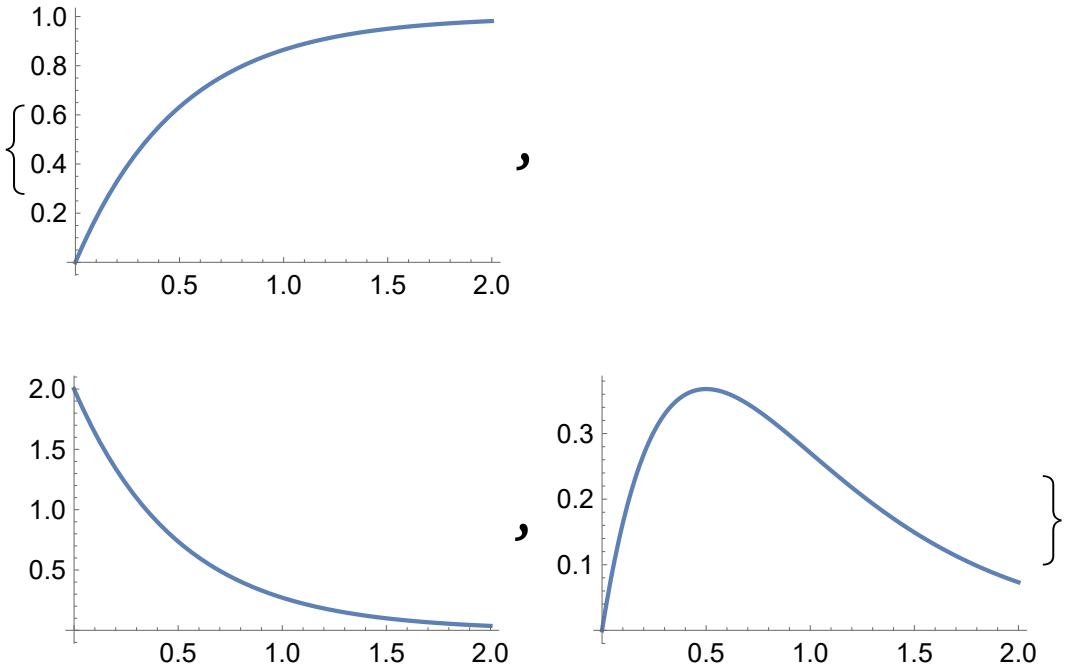
```
0.322926
```

```
0.333333
```

Relation to Exponential: Lorenz Curve

In order to build a Lorenz curve, we will need the inverse of the CDF. This is possible because our CDFs are strictly increasing.

```
{Plot[CDF[ExponentialDistribution[2]][x],
  {x, 0, 2}],
 Plot[PDF[ExponentialDistribution[2]][x],
  {x, 0, 2}],
 Plot[x * PDF[ExponentialDistribution[2]][x],
  {x, 0, 2}]}
```




```
Clear[x0, x, α]
cdfExp = Simplify[
   CDF[ExponentialDistribution[α]][x], x ≥ 0]
Simplify[Solve[s == cdfExp, x, Reals],
 1 > s ≥ 0 && α > 1] (* invert the CDF *)
```

```
pdfExp = Simplify[
   PDF[ExponentialDistribution[α]][x], x ≥ 0]
meanExp = Assuming[x0 > 0 && α > 1,
   Integrate[x * pdfExp, {x, 0, +Infinity}]
 ]
Assuming[x0 > 0 && α > 1,
 Integrate[x * pdfExp, {x, 0, t}] / meanExp
]
% /. {t → -Log[1 - s] / α}
Plot[{%, s^3}, {s, 0, 1},
 (* plot s^3 just for comparison *)
 AspectRatio → 1,
 PlotStyle → {Red, Gray}]
```

$$1 - e^{-x\,\alpha}$$

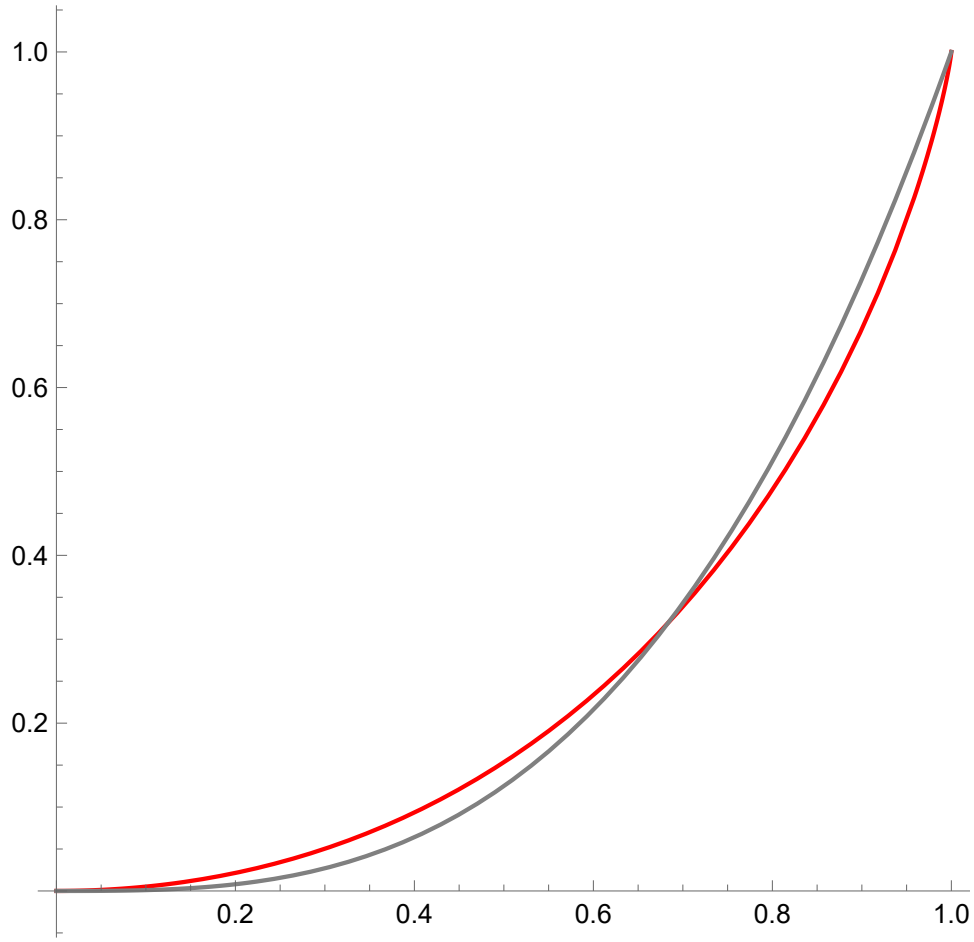
$$\left\{\left\{x \to -\frac{\text{Log}[1-s]}{\alpha}\right\}\right\}$$

$$e^{-x\,\alpha}\,\alpha$$

$$\frac{1}{\alpha}$$

$$1 - e^{-t\,\alpha}\,(1 + t\,\alpha)$$

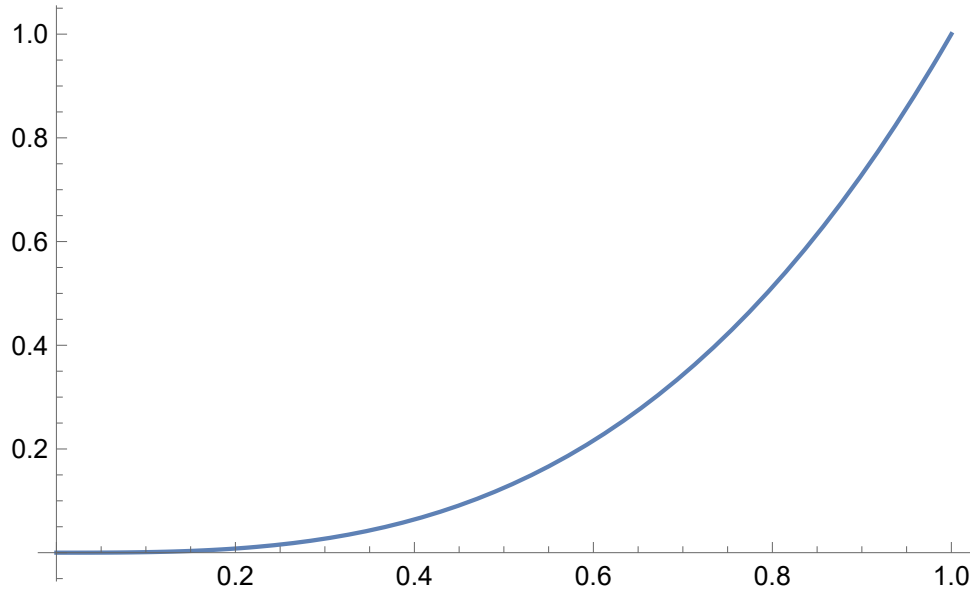
$$1 - (1-s)\,(1 - \text{Log}[1-s])$$


1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0

Yunker's Lorenz Curve

Yunker expresses the Lorenz curve as

```
yshareYunker[pshare_, g_] := pshare^g
Plot[yshareYunker[s, 3.0], {s, 0, 1}]
yunkerB = Assuming[g > 1,
  Integrate[yshareYunker[s, g], {s, 0, 1}]]
yunkerGini = (1 / 2 - yunkerB) / (1 / 2) // Simplify
(1 + yunkerGini) / (1 - yunkerGini) // Simplify
(* compute incremental income share *)
Clear[n]
Simplify[yshareYunker[(1 + i) / n] -
  yshareYunker[i / n], 0 < i < n]
```



$$\frac{1}{1+g}$$

$$\frac{-1+g}{1+g}$$

```
g
```

$$-\left(\frac{i}{n}\right)^{g}+\left(\frac{1+i}{n}\right)^{g}$$

Gini Coefficient

The Gini Coefficient is twice the area between the 45 degree line and the Lorenz curve. We can calculate that area as

```
Assuming[α > 1,
 Integrate[s - 1 + (1 - s)^(1 - 1/α), {s, 0, 1}]
]
gini = 2 * % // Simplify
```

$$\frac{1}{-2+4\,\alpha}$$

$$\frac{1}{-1+2\,\alpha}$$

```
Solve[g == gini, α]
```

$$\left\{\left\{\alpha \to \frac{1+g}{2\,g}\right\}\right\}$$

```
Solve[δ == 1 - 1 / α, α] // Flatten
gini /. % // Simplify
Solve[g == %, δ]
```

$$\left\{\alpha \to \frac{1}{1-\delta}\right\}$$

$$\frac{1-\delta}{1+\delta}$$

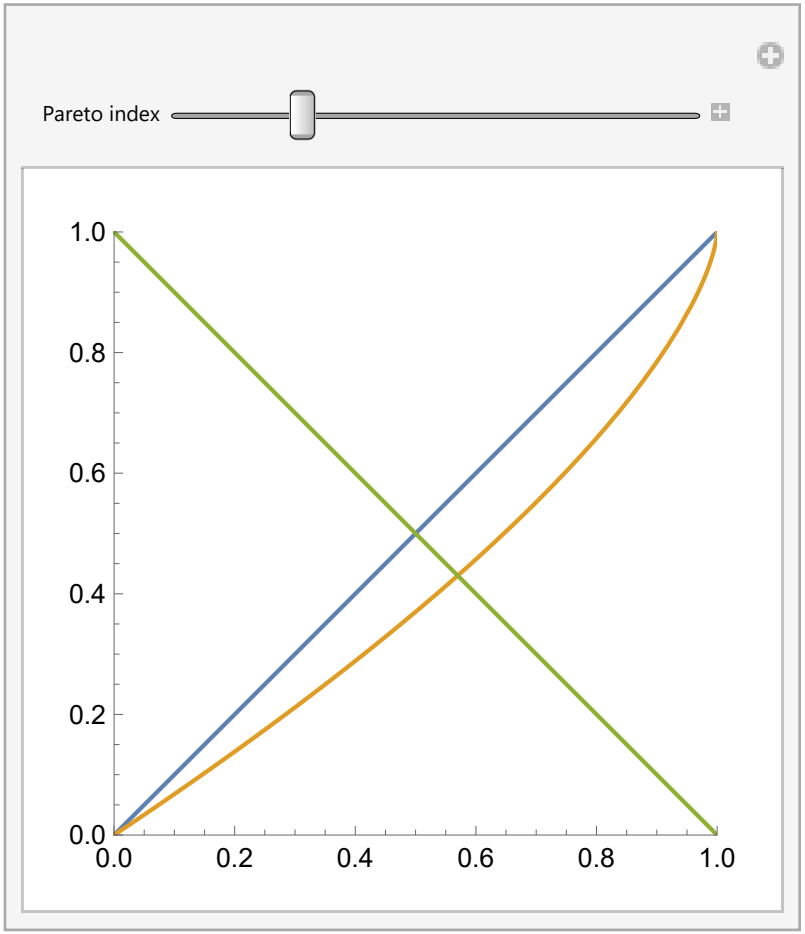
$$\left\{\left\{\delta \to \frac{1-g}{1+g}\right\}\right\}$$

Pareto Distribution and Lorenz Curve

In[85]:=

```
L[F_, k_] := 1 - (1 - F)^(1 - 1./k);
options03 = {PlotRange → {{0, 1}, {0, 1}},
   AspectRatio → 1, ImageSize → 250};
Manipulate[Plot[{F, L[F, k], 1 - F},
   {F, 0, 1}, Evaluate[options03]],
 {{k, 3, "Pareto index"}, 1, 10}]
```

Out[86]=



80-20 Rule: Pareto (1906) noticed that about 80% of the land in Italy was owned by about 20% of the

population.

However his British tax return data showed something closer to 70-30.

There will always be some such proportion: look for where the Lorenz curve crosses the unit simplex.

80-20 Rule

We have seen with $\alpha = 2$ that 1% of the population has a size at least 10 times the minimum, and 1% of that 1% has a size 10 times that.

More generally, if $\alpha > 1$ (so that the expected value is finite), is some fraction $0 \le f \le 1/2$ such that f of those sampled receive $(1 - f)$ of all income, and similarly for every real (not necessarily integer) n > 0, 100pn % of all people receive 100(1 – p)n % of all income.

```
Assuming[α > 1,
 Solve[1 - s == 1 - (1 - s)^(1-1/α), s]
]
Assuming[1 > δ > 0,
 Solve[s == (1 - s)^δ, s]
]
Assuming[δ > 0,
 Solve[s + s^(1/δ) == 1, s]
]
```

Solve::nsmet : This system cannot be solved with the methods available to Solve.

>>

```
Solve[1 - s == 1 - (1 - s)^(1 - 1/α), s]
```

Solve::nsmet : This system cannot be solved with the methods available to Solve.

>>

```
Solve[s == (1 - s)^δ, s]
```

Solve::nsmet : This system cannot be solved with the methods available to Solve.

>>

```
Solve[s + s^(1/δ) == 1, s]
```

```
Solve[1 - s == 1 - (1 - s)^(1 - 1/α), α]
```

Solve::ifun : Inverse functions are being used by Solve, so some solutions may not be found; use Reduce for complete solution information. >>

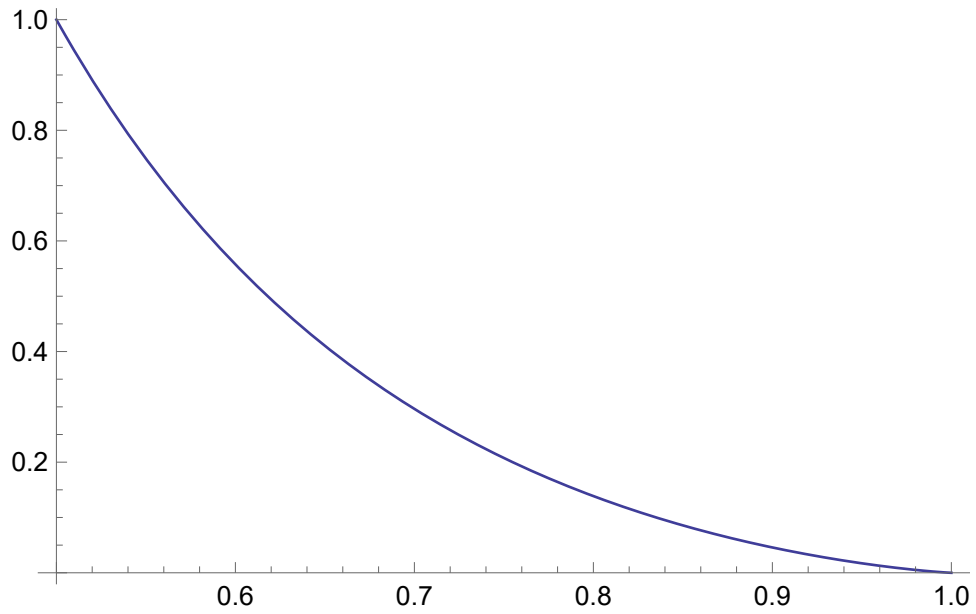
$$\left\{\left\{\alpha \to \frac{\text{Log}[1 - s]}{\text{Log}[1 - s] - \text{Log}[s]}\right\}\right\}$$

```
Solve[s + s^(1/δ) == 1, δ]
```

Solve::ifun : Inverse functions are being used by Solve, so some solutions may not be found; use Reduce for complete solution information. >>

$$\left\{\left\{\delta \to \frac{\text{Log}[s]}{\text{Log}[1 - s]}\right\}\right\}$$

```
Plot[Log[s] / Log[1 - s], {s, 0.5, 1}]
```



Data Creation and Analysis

Sampling from Power Law (Pareto) Distributions

We can generate a sample from a Pareto distribution by sampling from a uniform distribution on (0,1].
We transform each point U from the uniform according to $X = x_0 \,/\, U^{1/\alpha}$.

Then looking at the survival function we have

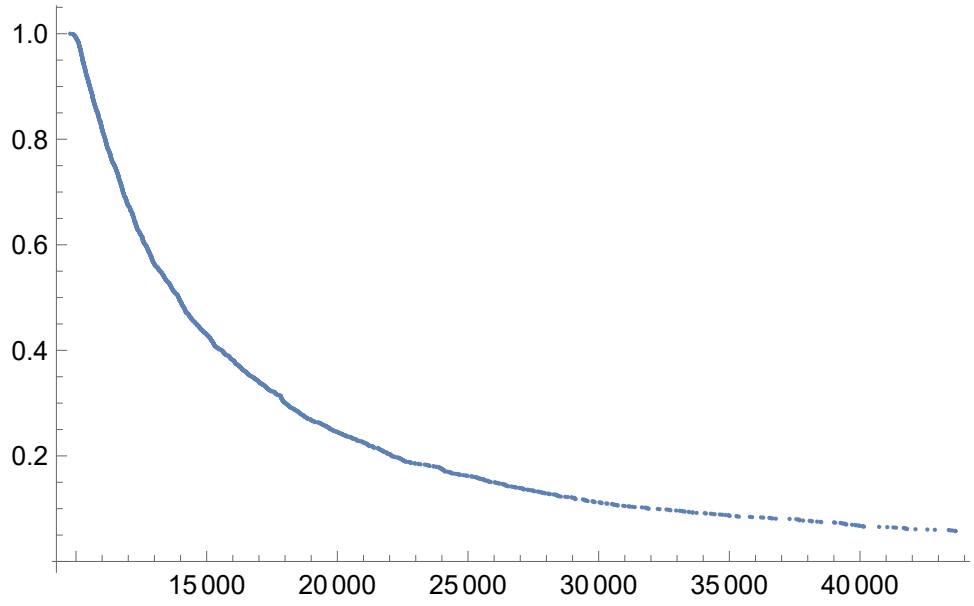
$$P[X > x] = P\left[\text{x0} \,/\, U^{1/\alpha} > x\right] =$$

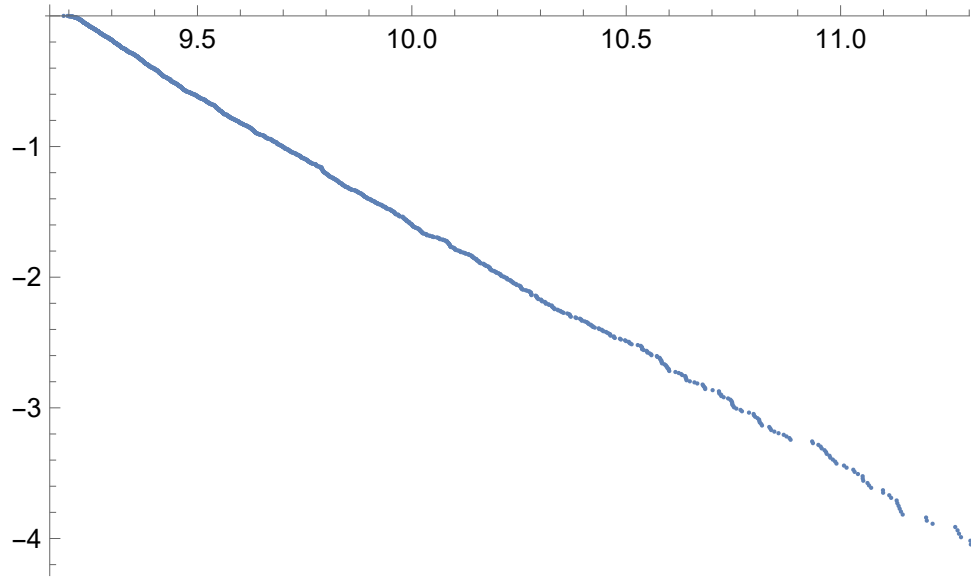
$$P\left[\, U^{1/\alpha} < \text{x0}/x\right] = P[U < (\text{x0}/x)^{\alpha}] = (\text{x0}/x)^{\alpha}$$

Technical note: note that we must rule out drawing a

0 from our uniform distribution. Most software draws from the interval [0, 1) . In this case, just use 1 – *U* for your sample.

```
Clear[sizedata, incomes]
alpha = 2; xmin = 10 000; npts = 2000;
sizedata =
  xmin / (1 - RandomReal[1, npts]) ^ (1 / alpha);
noise = RandomVariate[
   NormalDistribution[0, 100], npts];
sizedata = Sort[sizedata + noise];
Clear[noise]
proportionlarger =
  Reverse[Range[npts]] / npts // N;
survivaldata =
  Transpose[{sizedata, proportionlarger}] // N;
ListPlot[survivaldata]
llplot = ListPlot[Log[survivaldata]]
```

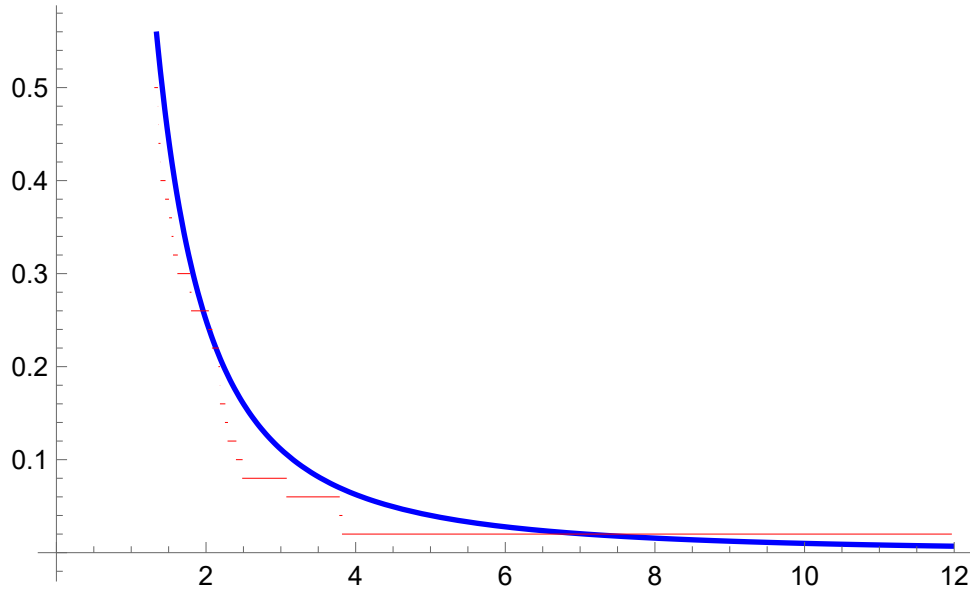

1.0
0.8
0.6
0.4
0.2
15 000
20 000
25 000
30 000
35 000
40 000


9.5
10.0
10.5
11.0
−1
−2
−3
−4

Pareto Distribution (Empirical Tail Function)

Large values occur with low frequency. Due to sample-size limitations, the empirical survival function will generally be zero for many values. Here we compare the theoretical distribution with an empirical distribution.

```
Clear[x0, α, x];
data =
  RandomVariate[ParetoDistribution[1, 2], 50] ;
(* x0=1; α=2 *)
maxx = Max[data];
𝒟 = EmpiricalDistribution[data];
dataPlotSFP = Plot[SurvivalFunction[𝒟, x],
   {x, 0, maxx}, PlotStyle → {Thin, Red}];
theoryPlotSFP = Plot[
   SurvivalFunction[ParetoDistribution[1, 2], x],
   {x, 0, maxx}, PlotStyle → {Thick, Blue}];
Show[theoryPlotSFP, dataPlotSFP]
```


0.5
0.4
0.3
0.2
0.1
2
4
6
8
10
12

Simplest Approach to Estimation

An obvious estimator for x_0 is the minimum observation (which is also the maximum likelihood estimator). Recall that the mean of the distribution is $x_0\,\alpha/(\alpha-1)$, we can then estimate α using

```
Clear[mean, x0, α]
minsize = Min[sizedata]
{meansize = Mean[sizedata],
 theoreticalmean = xmin * alpha / (alpha - 1)}
Solve[mean == x0 * α / (α - 1), α] /.
 {mean → meansize, x0 → minsize}
```

```
9776.72
```

```
{20 242.6, 20 000}
```

```
{{α → 1.93415}}
```

Not bad.

Method of Moments

We might improve a little on this by estimating x_0 with using the expected value for the minimum observation given the sample size. (See http://www.math.umt.edu/gideon/pareto.pdf) Let us adjust the minimum to account for sample size. (We have a large sample, so it does not make much difference.)

```
eq1 = Mean[sizedata] == k * α / (α - 1)
eq2 = Min[sizedata] == k * n * α / (n * α - 1) /.
   {n → Length[sizedata]}
Solve[eq1 && eq2, {k, α}]
```

$$20\,242.6 == \frac{k\,\alpha}{-1+\alpha}$$

$$9776.72 == \frac{2000\,k\,\alpha}{-1+2000\,\alpha}$$

Solve::ratnz : Solve was unable to solve the system with inexact coefficients. The answer was obtained by solving a corresponding exact system and numericizing the result. >>

$$\{\{k \to 9774.2,\ \alpha \to 1.93368\}\}$$

Maximum Likelihood

The “likelihood” of seeing this data for given values of x_0 α and α is given by the joint density function. Assuming independent draws, this is the product of the PDF values for each observation:

$$L(x_0, \alpha; x) = \prod_{i=1}^{n} \alpha \frac{x_0^{\alpha}}{x_i^{\alpha+1}} = \prod_{i=1}^{n} \frac{\alpha}{x_0} \frac{x_0^{\alpha+1}}{x_i^{\alpha+1}}$$

We can make this likelihood as big as we wish by increasing x_0, but we are constrained by the smallest value actually observed. The maximum likelihood

estimate of x_0 is therefore the minimum realized value. In order to find the maximum likelihood estimate of α, we will work with the logarithm of the likelihood:

$$\text{llf}(x_0, \alpha; x) = n \,\text{Log}[\alpha / x_0] - \sum_{i=1}^{n} (\alpha + 1) \,\text{Log}[x_i / x_0]$$

The slope of this (with respect to α) is

$$\frac{n}{\alpha} - \sum_{i=1}^{n} \text{Log}[x_i / x_0]$$

so the slope is 0 when

$$\alpha = n / \sum_{i=1}^{n} \text{Log}[x_i / x_0]$$

```
Length[sizedata] /
 Total[Log[sizedata / Min[sizedata]]]
```

```
1.92451
```

Fitting a Pareto Distribution to the Data

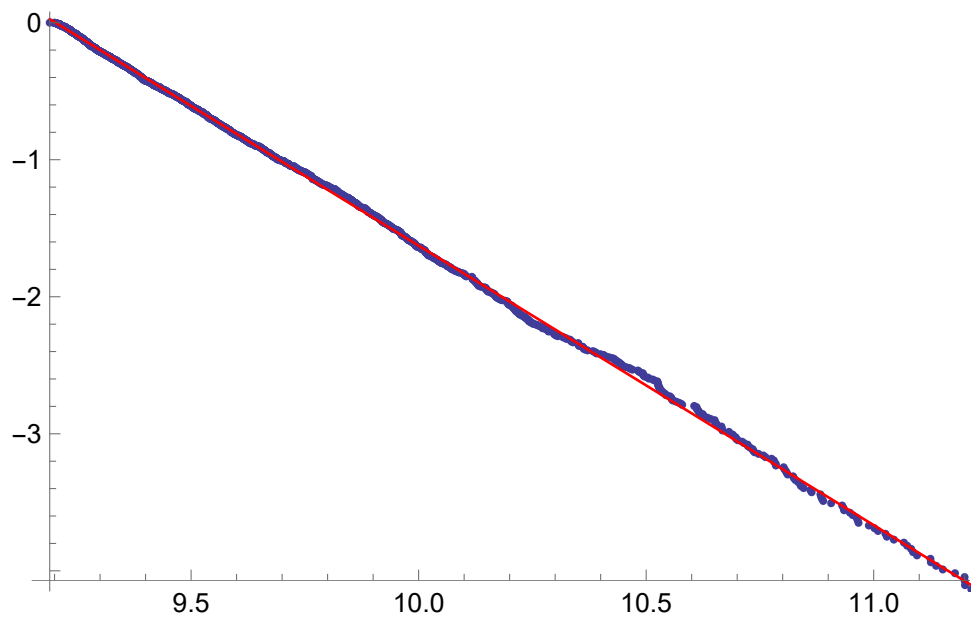
Recall that the survivial function told us that the proportion surviving is linear in log(x0/*x*). So we can look for a simple linear fit. The coefficient on *x* is our estimate of the Pareto index.

```
alpha
fit = Fit[Log[survivaldata], {1, x}, x]
(* linear fit to logged data *)
coefs = CoefficientList[fit, x]
Show[
 {llplot, Plot[fit, {x, Log[First[sizedata]],
    Log[Last[sizedata]]}, PlotStyle → {Red}]}]
Exp[-coefs[[1]] / coefs[[2]]]
(* implied value of x0 *)
```

```
2
```

```
18.7679 - 2.03951 x
```

```
{18.7679, -2.03951}
```




```
9918.88
```

A nonlinear model fit of the survival function produces similar results.

```
nlm01 = NonlinearModelFit[survivaldata,
  SurvivalFunction[ParetoDistribution[
    x0hat, alphahat], $x], {x0hat, alphahat}, $x]
```

FittedModel

$$\left\{\begin{array}{ll} \dfrac{9.66847\times10^{7}}{\$x^{\ll 19\gg}} & \$x \geq 9949.45 \\ 1 & \text{True} \end{array}\right.$$

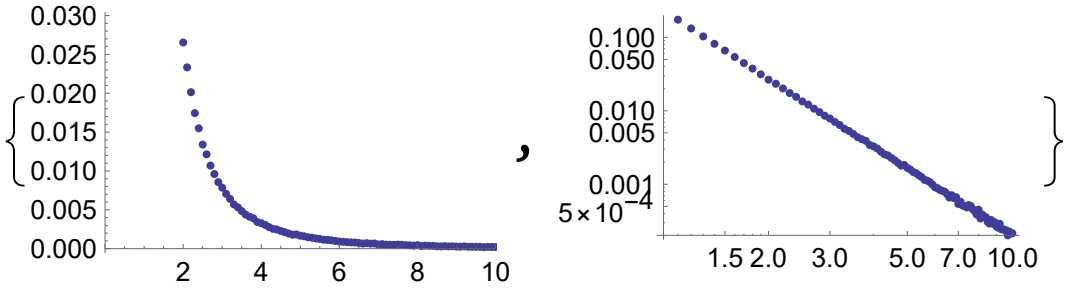
Estimation Considerations

Gabaix and Ibragimov (2011) note that in small samples OLS estimates of α are downward biased and the conventional standard error are too small.

Sampling from Power Law (Pareto) Distributions

We are often forced to work with binned data. Let's create some.

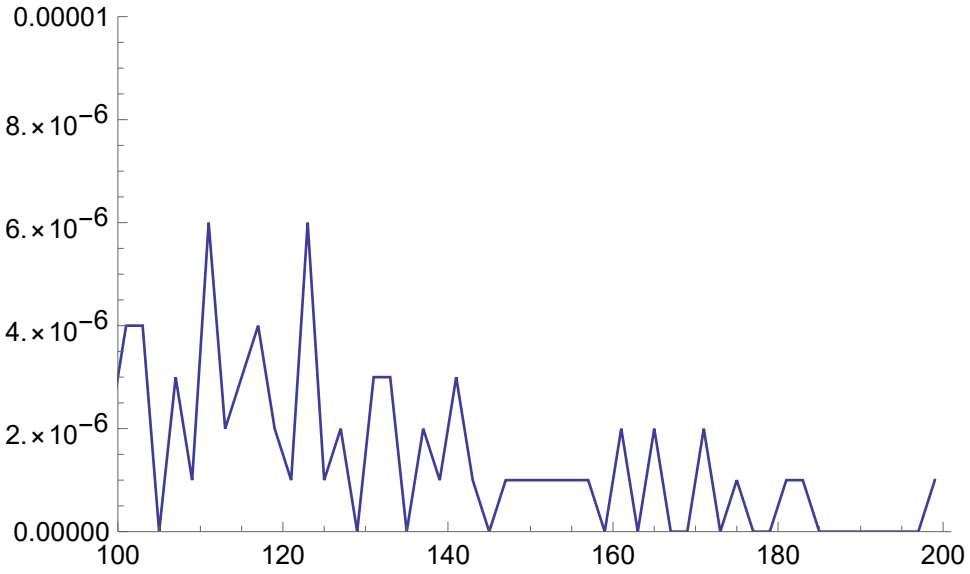
```
npts = 10^6;
alpha = 2;
x0 = 1;
x = x0 / (1 - RandomReal[1, npts])^(1 / alpha);
(* we'll need x again for next figure *)
xmax = 10 * x0;
bins = Table[i, {i, x0, xmax, xmax / 100}];
relfreq = BinCounts[x, {bins}] / npts;
{ListPlot[Transpose[{bins[[2 ;;]], relfreq}],
  PlotRange → {{0, xmax}, Automatic}],
 ListLogLogPlot[
  Transpose[{bins[[2 ;;]], relfreq}]]}
```


0.030
0.025
0.020
0.015
0.010
0.005
0.000
2
4
6
8
10
,
0.100
0.050
0.010
0.005
0.001
5×10^−4
1.5
2.0
3.0
5.0
7.0
10.0

Power Law Frequency Distribution

In our last slide, we cheated a bit by showing only the bins for relatively small sizes, which occur with the greatest frequency. As size increases, relative frequency falls, and statisitcal noise becomes more prominent, even if we substantially increase bin size.

```
xmax = 200 * x0;
bins = Table[i, {i, x0, xmax, xmax / 100}];
relfreq = BinCounts[x, {bins}] / npts;
mypoints = Transpose[{Rest[bins], relfreq}];
ListLinePlot[mypoints,
 PlotRange → {{100 * x0, Automatic}, {0, 10^-5}}]
```



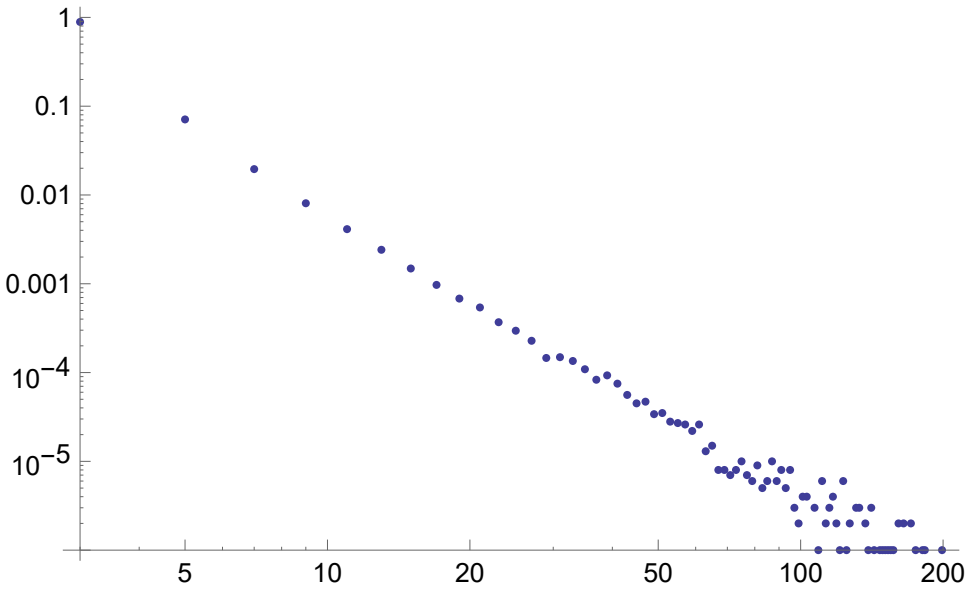
```
{{x}} /. {{x} → {x, y}, x → 1}
```

```
{{x, y}}
```

Power Law: Log Scale

We might hope to address this by moving to a log scale. This proves informative but is only partially successful. Why? Our bins are still linear.
Notice the empty bins for large sizes.

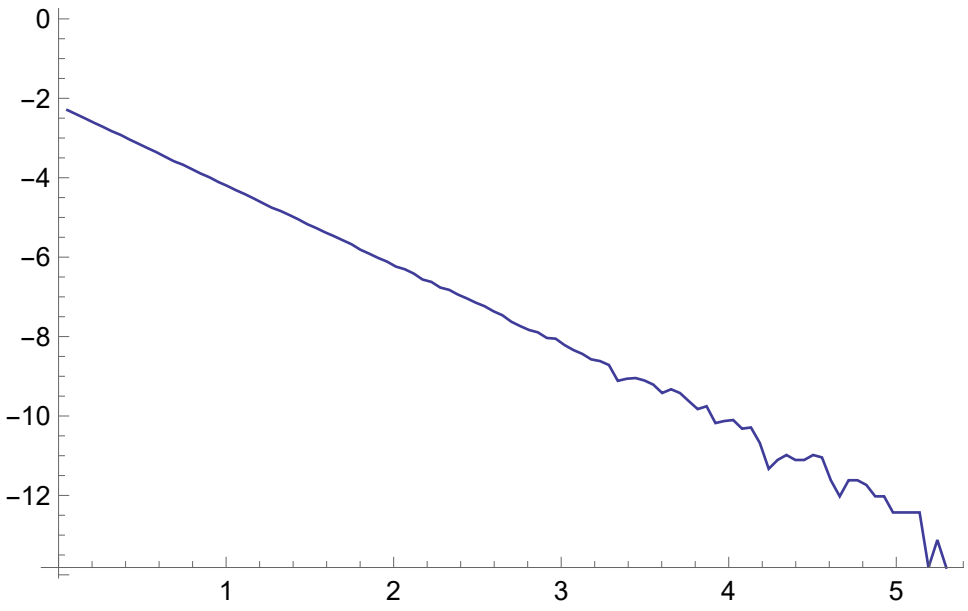
```
ListLogLogPlot[mypoints]
```



Logarithmic Binning

It works better to let our bin size grow as we consider larger size realizations: we can use logarithmic binning.

```
logxmax = Log[xmax];
logbins = Table[i, {i, 0, logxmax, logxmax / 100}];
relfreq = BinCounts[Log[x], {logbins}] / npts;
data =
  Transpose[{Rest[logbins], Log[relfreq]}];
ListLinePlot[data]
```




```
Clear[x]
```

Nonlinear Curve Fitting

Some Census Data

Data from http://www.census.gov/hhes/www/income/data/historical/inequality/IE-1.pdf table A.3. Income is in 2010 dollars.

Measures of income dispersion	2010	2009[1]	2008	2007	2006	2005	2004[2]	2003	2002	2001	2000[3]
MEASURE											
Household Income at Selected Percentiles											
10th percentile limit	11,904	12,320	12,315	12,789	12,977	12,607	12,589	12,490	12,871	13,160	13,398
20th percentile limit	20,000	20,791	20,974	21,337	21,666	21,419	21,338	21,320	21,713	22,131	22,689
50th (median)	49,445	50,599	50,939	52,823	52,124	51,739	51,174	51,353	51,398	52,005	53,164
80th percentile limit	100,065	101,651	101,508	105,156	104,930	102,420	101,580	102,980	101,824	102,833	103,525
90th percentile limit	138,923	139,904	140,050	143,012	143,825	140,823	139,514	140,125	138,299	139,937	141,805
95th percentile limit	180,810	182,972	182,277	186,126	188,175	185,397	181,399	182,707	181,797	185,345	183,865

Census Data

```
h01ar = SemanticImport[
   "D:/data/census/2018-tableH01AR.csv"]
```

Year	kCount	Percentile20	Percentile40	Percentile60	Percentile80	Percentile95
2016	126224	24002	45600	74869	121018	225251
2015	125819	23088	44061	72911	118480	217172
2014	124587	21728	41754	69153	113811	209419
2013	123931	21638	42282	69242	113582	211362
2013	122952	21535	41408	67492	109129	201957
2012	122459	21533	41568	67511	108818	199827
2011	121084	21617	41096	66609	108375	198438
2010	119927	22017	41832	67702	110116	198686
2009	117538	22880	43124	69134	111865	201359

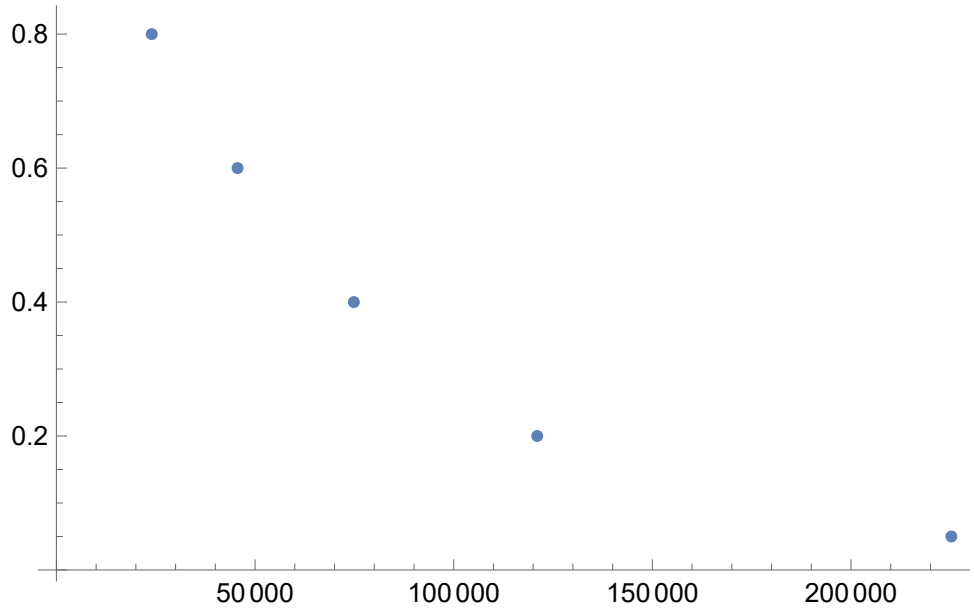
2009	117 538	22 880	45 124	69 134	111 865	201 359
2008	117 181	23 089	43 476	69 924	111 744	200 658
2007	116 783	23 489	45 262	71 770	115 758	204 892
2006	116 011	23 850	44 967	71 425	115 508	207 146
2005	114 384	23 570	44 244	70 864	112 705	204 014
2004	113 343	23 489	44 059	70 177	111 818	199 682
2003	112 000	23 468	44 369	71 059	113 358	201 120
2002	111 278	23 911	44 545	70 950	112 127	200 192
2001	109 297	24 361	45 162	71 849	113 195	204 021
2000	108 209	24 985	46 009	72 742	114 000	202 470
1999	106 434	24 702	46 014	72 630	114 216	204 698
1998	103 874	23 727	44 768	71 163	110 418	194 628

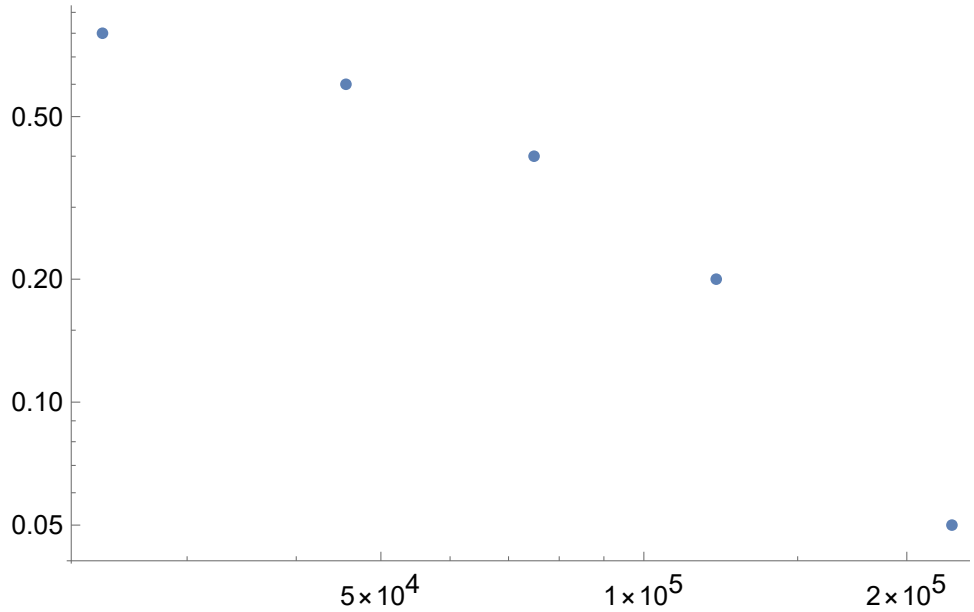
showing 1–20 of 51

Census Data

```
(* 2016 *)
pctls = {20, 40, 60, 80, 95} / 100;
cuts = {24 002, 45 600, 74 869, 121 018, 225 251};
survivals = 1 - pctls
ListPlot[{cuts, survivals}^T]
ListLogLogPlot[{cuts, survivals}^T]
```

$$\left\{\frac{4}{5}, \frac{3}{5}, \frac{2}{5}, \frac{1}{5}, \frac{1}{20}\right\}$$


0.8
0.6
0.4
0.2
50 000
100 000
150 000
200 000

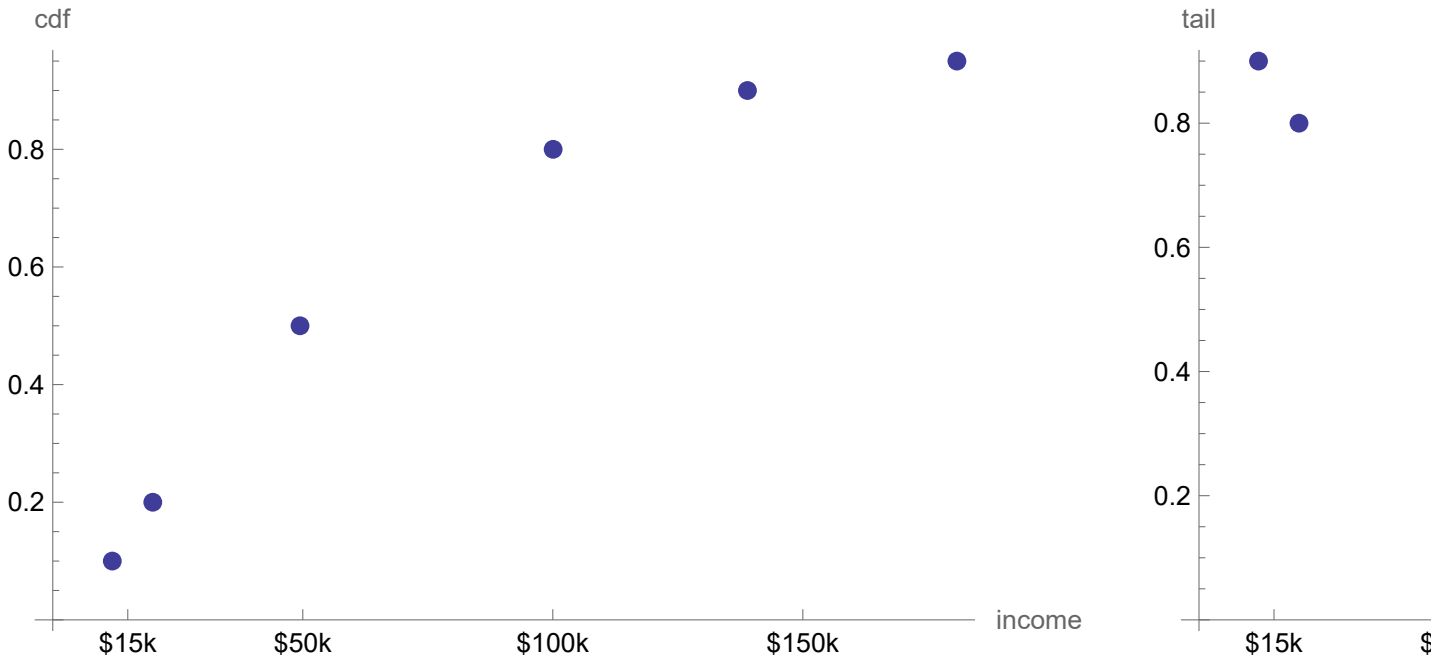

0.50
0.20
0.10
0.05
5×10^4
1×10^5
2×10^5

Census Data

```
incomes2010 =
  {11 904, 20 000, 49 445, 100 065, 138 923, 180 810};
cdf2010 = {10, 20, 50, 80, 90, 95} / 100.;

tail2010 = 1 - cdf2010;
incomecdf2010 =
  Transpose[{incomes2010, cdf2010}];
incometail2010 =
  Transpose[{incomes2010, tail2010}];
Labeled[GraphicsRow[{
```

```
g`cdf2010 = ListPlot[incomecdf2010, AxesLabel →
    {"income", "cdf"}, AxesOrigin → {0, 0},
   PlotStyle → PointSize[0.02],
   Ticks → {{{15 000, "$15k"}, {50 000, "$50k"},
      {100 000, "$100k"}, {150 000, "$150k"}},
     Automatic}, ImageSize → 400],
 g`tail2010 =
  ListPlot[incometail2010, AxesLabel →
    {"income", "tail"}, AxesOrigin → {0, 0},
   PlotStyle → PointSize[0.02],
   Ticks → {{{15 000, "$15k"}, {50 000, "$50k"},
      {100 000, "$100k"}, {150 000, "$150k"}},
     Automatic}, ImageSize → 400]
}], "2010 Census Data"]
```



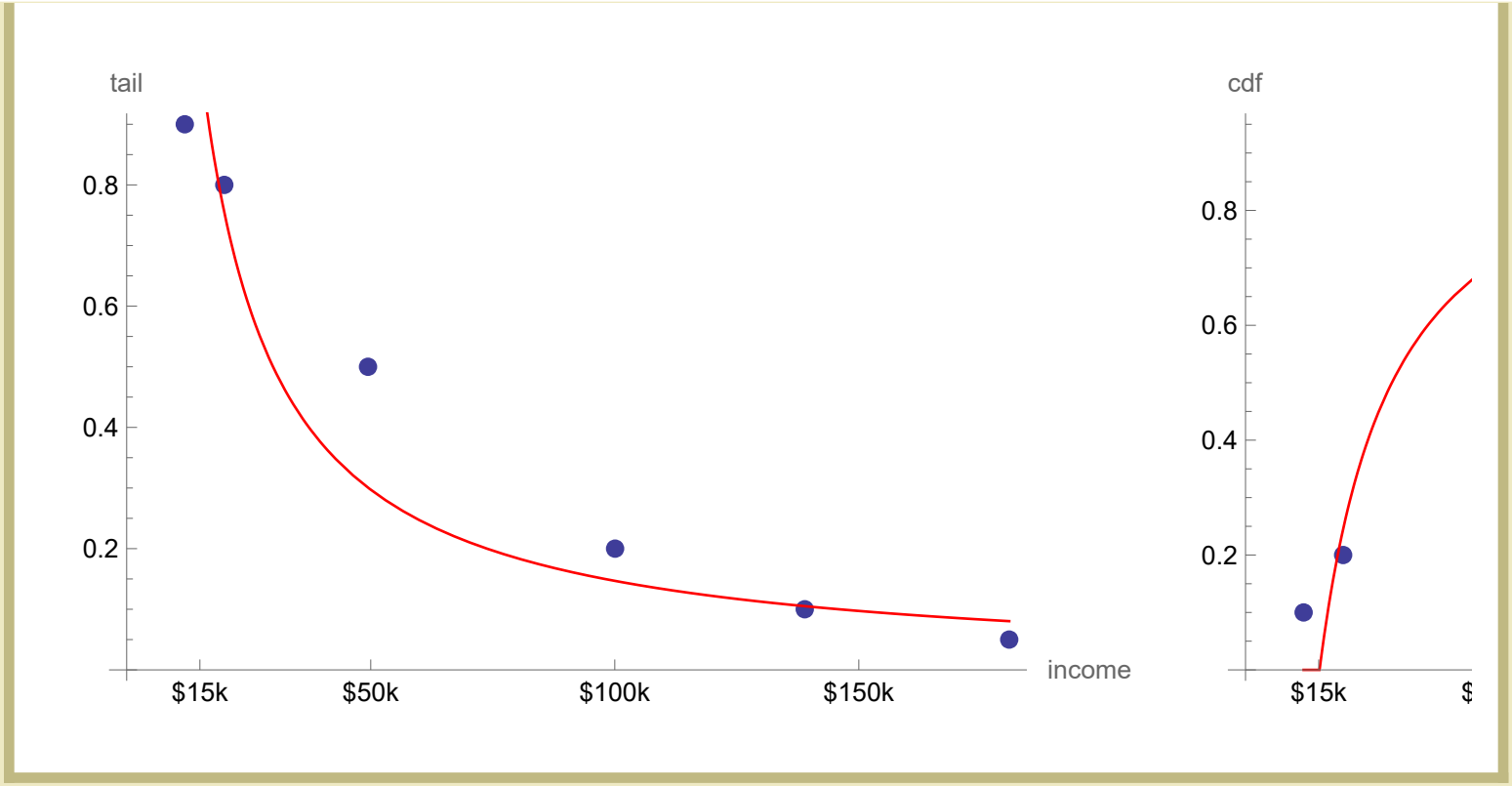
2010 Census Data

Loglinear Survival Fit

```
Clear[x]
fit2010 =
  Fit[Log[Transpose[{incomes2010, tail2010}]],
   {1, $x}, $x];
(* linear fit to logged data *)
coefs2010 = CoefficientList[fit2010, $x];
ahat2010 = -coefs2010[[2]];
x0hat2010 = Exp[coefs2010[[1]] / ahat2010 ];
(* implied value of x0 *)
tail2010fit = Piecewise[
```

```
   {{(x0hat2010 / x)^ahat2010, x > x0hat2010},
    {1, True}}]
cdf2010fit = 1 - tail2010fit;
GraphicsRow[{
  Show[{g`tail2010,
    Plot[tail2010fit, {x, First[incomes2010],
      Last[incomes2010]}, PlotStyle → {Red}]}]
  ,
  Show[{g`cdf2010,
    Plot[cdf2010fit, {x, First[incomes2010],
      Last[incomes2010]}, PlotStyle → {Red}]}]
 }]
```

$$\begin{cases} 17\,914.6\left(\frac{1}{x}\right)^{1.01727} & x > 15\,169.9 \\ 1 & \text{True} \end{cases}$$



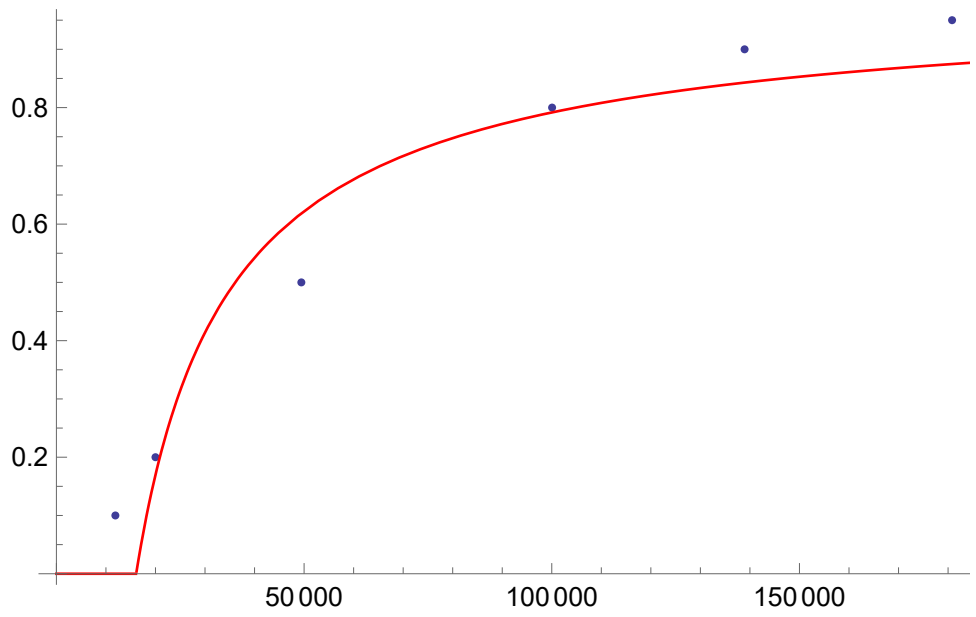
A problem with this log-linear survival fit it that it estimates minimum income at a value above the minimum observed value. But the same thing happens with a nonlinear estimation.

Nonlinear Least Squares: Fit CDF to 2010 Data

```
incomecdf2010 =
 {{11 904, 0.1}, {20 000, 0.2}, {49 445, 0.5},
  {100 065, 0.8}, {138 923, 0.9}, {180 810, 0.95}}
nlm01 = NonlinearModelFit[incomecdf2010,
   CDF[ParetoDistribution[khat, ahat], $x],
   {khat, ahat}, $x];
nlm01["BestFitParameters"]
gpareto2010 = Show[
  ListPlot[incomecdf2010, AxesOrigin → {0, 0}],
  Plot[nlm01[$x], {$x, 0, 200 000},
   PlotStyle → {Red}]]
```

```
{{11 904, 0.1}, {20 000, 0.2}, {49 445, 0.5},
 {100 065, 0.8}, {138 923, 0.9}, {180 810, 0.95}}
```

```
{khat → 16 104.6, ahat → 0.858956}
```



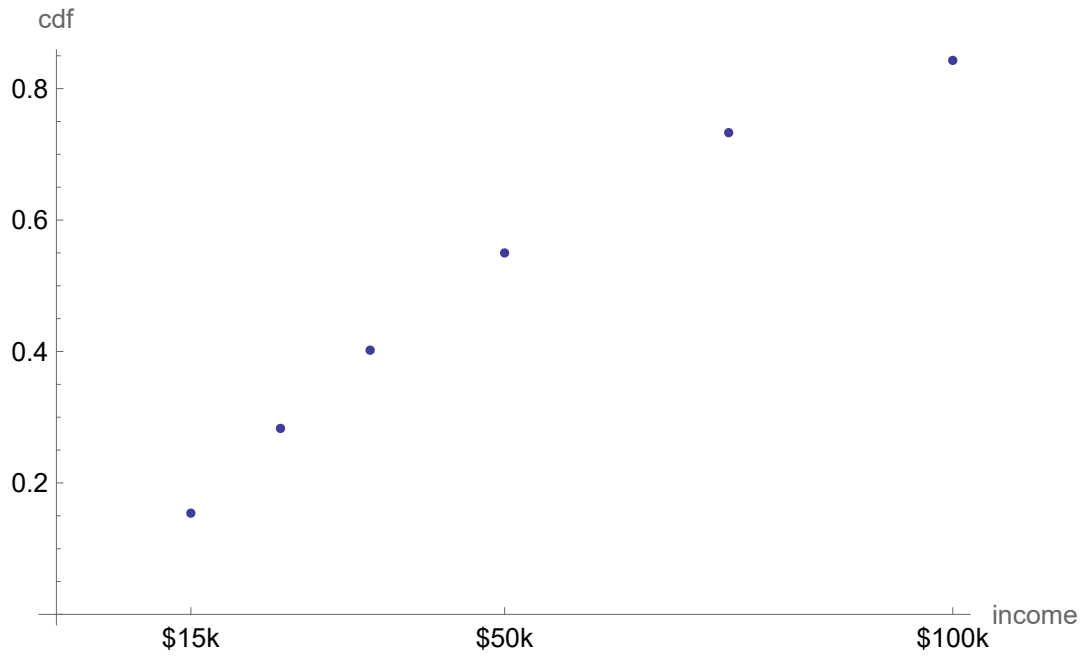
Some Earlier (2004) Income Data

For purposes of comparison, we use data from Maclachlan (2006).

```
incomes2004 =
  {15 000, 25 000, 35 000, 50 000, 75 000, 100 000};
cdf2004 =
  {0.154, 0.283, 0.402, 0.55, 0.733, 0.843};
tail2004 = 1 - cdf2004;
data2004fm = Transpose[{incomes2004, cdf2004}]
g`data2004fm = ListPlot[data2004fm,
  PlotStyle → PointSize[0.01], AxesOrigin → {0, 0},
  AxesLabel → {"income", "cdf"}, Ticks →
   {{{15 000, "$15k"}, {50 000, "$50k"}, {100 000,
```

```
      "$100k"}}, Automatic}, ImageSize → 400]
Clear[x]
fit2004 =
 Fit[Log[Transpose[{incomes2004, tail2004}]],
  {1, x}, x] (* linear fit to logged data *)
coefs2004 = CoefficientList[fit2004, x]
ahat2004 = -coefs2004[[2]]
x0hat2004 = Exp[coefs2004[[1]] / ahat2004 ]
(* implied value of x0 *)
cdf2004fit = 1 - (x0hat2004 / x) ^ahat2004
temp = Plot[cdf2004fit, {x, First[incomes2004],
    Last[incomes2004]}, PlotStyle → {Red}];
Show[{g`data2004fm, temp}]
```

```
{{15000, 0.154}, {25000, 0.283}, {35000, 0.402},
 {50000, 0.55}, {75000, 0.733}, {100000, 0.843}}
```



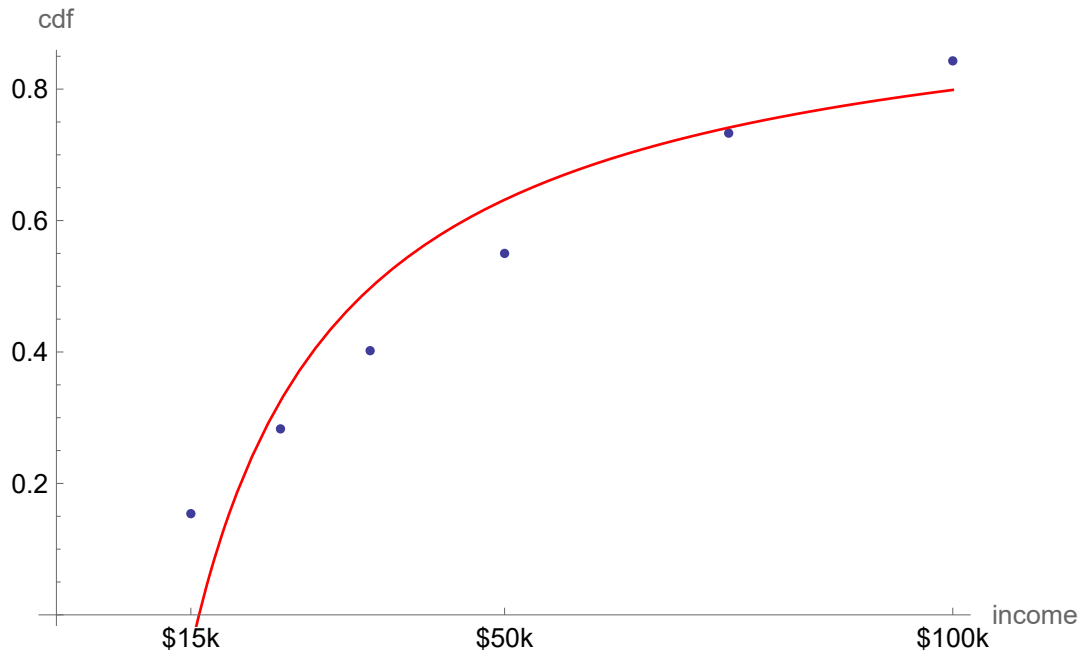

```
8.43993 - 0.872352 x
```

$$\{8.43993, -0.872352\}$$

$$0.872352$$

$$15\,913.4$$

$$1 - 4628.25 \left(\frac{1}{x}\right)^{0.872352}$$


cdf
0.8
0.6
0.4
0.2
$15k
$50k
$100k
income

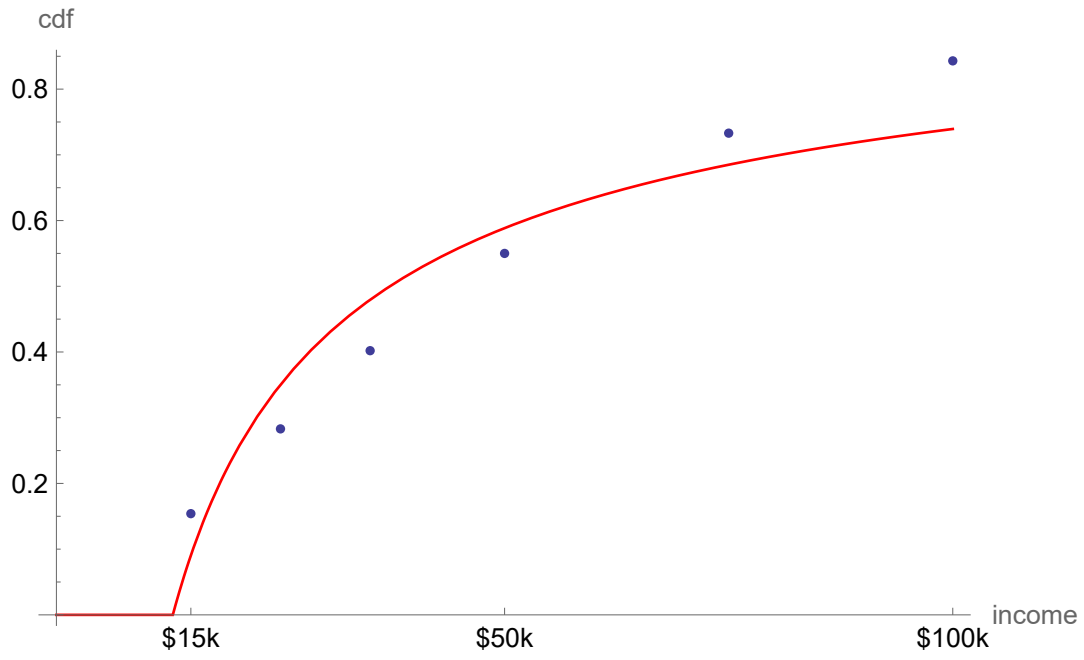
Fit to Pareto Distribution

Let’s fit these data points to a Pareto distribution, using NonlinearModelFit. (Mathematica 9 gives a perfect match to the same estimation on Maclachlan (2006), who used Mathematica 5.)

```
data2004fm ==
 {{15000, 0.154}, {25000, 0.283}, {35000, 0.402},
  {50000, 0.55}, {75000, 0.733}, {100000, 0.843}}
nlm01 = NonlinearModelFit[data2004fm,
   CDF[ParetoDistribution[khat, ahat], $x],
   {khat, ahat}, $x];
nlm01["BestFitParameters"]
g`pareto = Show[{g`data2004fm, Plot[nlm01[$x],
    {$x, 0, 100000}, PlotStyle → {Red}]}]
```

```
True
```

```
{khat → 12989.3, ahat → 0.658768}
```


cdf
0.8
0.6
0.4
0.2
$15k
$50k
$100k
income