The Math Behind the 80/20 Rule, Discrete Case

by Matt Borthwick

September, 14, 2017

Power-law Probability Distribution

Suppose that for each trial, the probability that the trial results in y instances of something (*e.g.* transactions, bugs, or distinct outputs) is distributed as a power law,

$$P\left(y \mid a\right) \quad \propto \quad y^{-a}$$

where y is an integer greater than zero. When $y = 0$, the power law fails, so instead let's generally set $P\left(y = 0\right) =$ some constant P_0 satisfying $0 \leq P_0 \leq 1$.

First we need to make sure this is a valid probability distribution. That is, we need all probabilities to sum to one. If we call the coefficient of proportionality Z, then

$$\begin{aligned} \sum_{y=0}^{\infty} P\left(y \mid a\right) &= 1 \\ P_0 + \sum_{y=1}^{\infty} Z y^{-a} &= 1 \end{aligned}$$

We can pull the Z outside of the sum. Inside the sum, y acts as a dummy variable, so let's change its name there to n; doing so will make our notation less confusing later.

$$\begin{aligned} P_0 + Z \sum_{n=1}^{\infty} n^{-a} &= 1 \\ Z &= \frac{1 - P_0}{\sum_{n=1}^{\infty} n^{-a}} \end{aligned}$$

We now know that our distribution will be valid if we use

$$P(y \mid a) = \begin{cases} P_0 & y = 0 \\ \left(\frac{1-P_0}{\sum_{n=1}^{\infty} n^{-a}}\right) y^{-a} & y = 1,\, 2,\, 3,\, \ldots \end{cases}$$

The particular form of our sum, $\sum_{n=1}^{\infty} n^{-a}$, is called the Riemann Zeta function and is usually denoted $\zeta(a)$. Unless a happens to be an even positive integer (or an odd negative integer), there's no simpler expression for $\zeta(a)$, but it is tabulated in various references or you can approximate it numerically. Discrete power-law distributions are thus commonly known as zeta distributions, or when the sum is over a finite range it's called the Zipf distribution. The continuous analogue is called the Pareto distribution.

Generalized Pareto Principle, Discrete Case

Now we can evaluate statements of the form, "f% of trials correspond to F% of instances." For example, "20% of lines of code generate 80% of the bugs," or "20% of the possible inputs to a function map to 80% of the possible outputs of the function."

To figure the f% of trials, we should include only the trials that give the highest results. We'll need to pick some appropriate starting value, y_f,

$$\begin{aligned} f &= P(y \geq y_f \mid a) \\ &= 1 - P(y < y_f \mid a) \\ &= 1 - P_0 - \sum_{y=1}^{y_f - 1} \frac{1-P_0}{\sum_{n=1}^{\infty} n^{-a}} y^{-a} \\ f &= (1-P_0)\left(1 - \frac{\sum_{y=1}^{y_f-1} y^{-a}}{\sum_{n=1}^{\infty} n^{-a}}\right) \end{aligned}$$

For the F% of resulting instances, we also start from y_f. Note that when counting the total number of instances, we need the sum over $yP(y \mid a)$, not just $P(y \mid a)$.

$$\begin{aligned} F &= \frac{\sum_{y=y_f}^{\infty} yP(y \mid a)}{\sum_{y'=0}^{\infty} y'P(y' \mid a)} \\ &= \frac{\sum_{y=y_f}^{\infty} y \frac{1-P_0}{\sum_{\ell=1}^{\infty} \ell^{-a}} y^{-a}}{0 + \sum_{m=1}^{\infty} m \frac{1-P_0}{\sum_{k=1}^{\infty} k^{-a}} m^{-a}} \end{aligned}$$

$$
\begin{aligned}
F &= \frac{\frac{1-P_0}{\sum_{\ell=1}^{\infty} \ell^{-a}} \sum_{y=y_f}^{\infty} y^{-(a-1)}}{\frac{1-P_0}{\sum_{k=1}^{\infty} k^{-a}} \sum_{m=1}^{\infty} m^{-(a-1)}} \\
&= \frac{\sum_{y=y_f}^{\infty} y^{-(a-1)}}{\sum_{m=1}^{\infty} m^{-(a-1)}} \\
&= \frac{\sum_{y=1}^{\infty} y^{-(a-1)} - \sum_{y=1}^{y_f-1} y^{-(a-1)}}{\sum_{m=1}^{\infty} m^{-(a-1)}} \\
F &= 1 - \frac{\sum_{y=1}^{y_f-1} y^{-(a-1)}}{\sum_{m=1}^{\infty} m^{-(a-1)}}
\end{aligned}
$$

This gives us two equations with two unknowns that we can solve numerically.

Incidentally, we can rewrite the last line above as

$$
\begin{aligned}
F &= 1 - \frac{\sum_{y=1}^{y_f-1} y^{-(a-1)}}{\sum_{m=1}^{\infty} m^{-(a-1)}} \\
&= 1 - \frac{P\left(0 < y < y_f \mid a-1\right)}{1-P_0}
\end{aligned}
$$

Because we started with F depending on $yP\left(y \mid a\right)$ and ended up with F depending on $P\left(y \mid a-1\right)$, you can begin to see how a zeta distribution with exponent a has a mean value that looks like a zeta distribution with exponent $a-1$. That's a convenient quirk of zeta and Pareto distributions, and we'll use it in the example below.

Specific 80/20 Example

What if we're given f and F and want to find an appropriate scaling exponent, a, and cutoff, y_f? Again, this almost always needs to be done numerically.

To replicate the situation in the "classic Pareto 80/20 rule", we want $P_0 = 0$, $f = 0.2$, and $F = 0.8$.

To generate those particular values, I found an approximate solution of $y_f = 3$ and $a \approx 2.15$.

Let's check that those numbers are correct. $\zeta\left(a = 2.15\right) = \sum_{n=1}^{\infty} n^{-2.15} \approx 1.5237$. When calculating F, we will also use $\zeta\left(a - 1 = 1.15\right) = \sum_{n=1}^{\infty} n^{-(2.15-1)} \approx 7.2547$.

y	$P(y \mid a = 2.15)$	$\frac{P(y\mid a=1.15)}{1-P_0}$
0	0	
1	$\frac{1^{-2.15}}{1.5237} \approx 65.63\%$	$\frac{1^{-1.15}}{7.2547} \approx 13.78\%$
2	$\frac{2^{-2.15}}{1.5237} \approx 14.78\%$	$\frac{2^{-1.15}}{7.2547} \approx 6.21\%$
3	$\frac{3^{-2.15}}{1.5237} \approx 6.18\%$	$\frac{3^{-1.15}}{7.2547} \approx 3.90\%$
4	$\frac{4^{-2.15}}{1.5237} \approx 3.33\%$	$\frac{4^{-1.15}}{7.2547} \approx 2.80\%$
5	$\frac{5^{-2.15}}{1.5237} \approx 2.06\%$	$\frac{5^{-1.15}}{7.2547} \approx 2.17\%$
6	$\frac{6^{-2.15}}{1.5237} \approx 1.39\%$	$\frac{6^{-1.15}}{7.2547} \approx 1.76\%$
7	$\frac{7^{-2.15}}{1.5237} \approx 1.00\%$	$\frac{7^{-1.15}}{7.2547} \approx 1.47\%$
8	$\frac{8^{-2.15}}{1.5237} \approx 0.75\%$	$\frac{8^{-1.15}}{7.2547} \approx 1.26\%$
9	$\frac{9^{-2.15}}{1.5237} \approx 0.58\%$	$\frac{9^{-1.15}}{7.2547} \approx 1.10\%$
10	$\frac{10^{-2.15}}{1.5237} \approx 0.46\%$	$\frac{10^{-1.15}}{7.2547} \approx 0.98\%$
$\vdots$	$\vdots$	$\vdots$

You can read row 7 in this table as, "1.00% of our clients made exactly 7 transactions, and together, those clients accounted for 1.47% of all transactions."

At first glance, it might seem simplest to get f by summing the entries in rows 3 through 10 of the middle column. However, the values of rows 11 and higher are significant too. In fact, you might have to sum hundreds, thousands, or even millions of rows before the sum converges enough.

Instead, it's far more efficient to use the second line of our formula for f above,

$$\begin{aligned} f &= 1 - P(y < y_f \mid a) \\ &= 1 - P(y < 3 \mid a = 2.15) \\ &= 100\% - 0 - 65.63\% - 14.78\% \\ &= 19.59\% \end{aligned}$$

or approximately 20%. Similarly,

$$\begin{aligned} F &= 1 - \frac{P(0 < y < y_f \mid a - 1)}{1 - P_0} \\ &= 1 - \frac{P(0 < y < 3 \mid a = 1.15)}{1 - 0} \\ &= 100\% - \frac{13.78\% + 6.21\%}{1} \\ &= 80.00\% \end{aligned}$$

A More Complicated Example

It's just a coincidence of the 80/20 assumption that $F = 1 - f$. This coincidence allows us to flip things backwards and think of 20% of F as generating 80% of f.

As a different example without such a coincidence, I encourage you to try playing around with $P_0 = 52\%$, $f = 5\%$, and $F = 58\%$. (Clearly $F \neq 1 - f$ in this case.)

You should find that $y_f = 4$ and $a \approx 2.30$ works. Here's row 7 so you can check your work:

y	$P(y \mid a = 2.30)$	$\frac{P(y\|a=1.30)}{1-P_0}$
0	52%	
⋮	⋮	⋮
7	0.38%	2.03%
⋮	⋮	⋮

Two Other Things I Should Mention

Working with arbitrarily chosen numbers like these is not particularly useful in real life. Instead, people usually take measured data, $\hat{P}(y)$, and do a linear regression of $\ln \hat{P}(y)$ vs. $\ln y$. The best-fit slope parameter returned by the regression is also the best fit to our scaling exponent, a.

By the way, when $a < 3$ (as it is in our example above), the variance of the zeta and Pareto distributions is infinite. This means the "long tail" of the distribution will persist in spite of any attempts to apply the Central Limit Theorem. And even when $a \geq 3$, it may require aggregating a very large number of measurements for the "tail" to decay away as quickly as one expects for a normal distribution.