Member-only story

Pareto, Power Laws, and Fat Tails

What they don't teach you in statistics

Shaw Talebi

Published in Towards Data Science · 12 min read · Nov 11, 2023

A Black Swan. Image from Canva.

Statistics is the bedrock of data science and analytics. It gives us a powerful toolbox to objectively answer complex questions. However, many of our

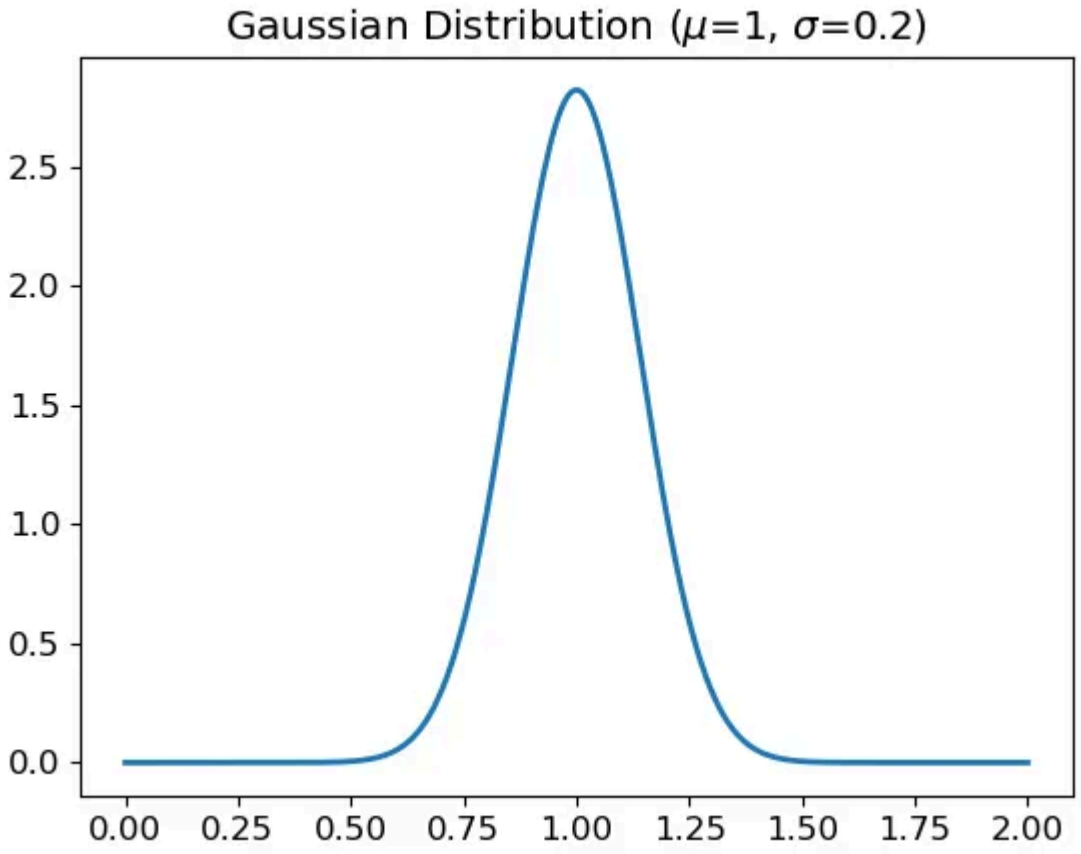
favorite statistical tools become useless when applied to a particular class of data — Power Laws.

In this article, I will provide a beginner-friendly guide to Power Laws and describe 3 major problems with using traditional statistical methods to analyze them.

Table of Contents

Weighing Your Barista

Many quantities in nature tend to clump around a typical value. For example, if you sat in a (busy) coffee shop and measured the weights of all the baristas and customers going in and out, you would (eventually) observe a pattern like the plot below.



Example Gaussian distribution. Technical Note: when measuring adult human weight, a Gaussian-like distribution will appear for each sex. Image by author.

This plot is an example of a **Gaussian distribution**, which you may have encountered in STAT 101 or business statistics. The beauty of a Gaussian is that we can **capture much of the essential information** of the underlying thing (e.g. weights of baristas) **with just a single number** — the **mean.**

Going further, we can get even more information by characterizing ***how spread out*** the data are via measures like **standard deviation** and **variance**.

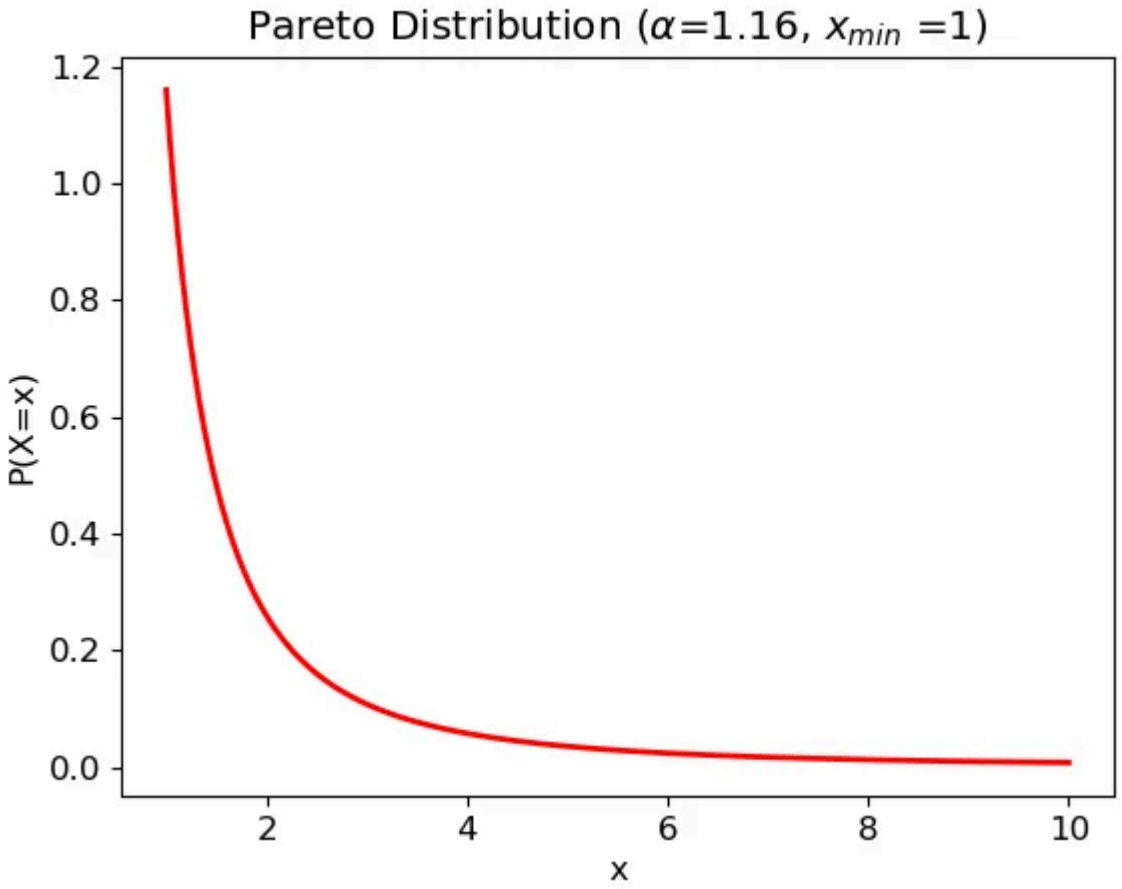
These concepts everyone learns in introductory statistics give us a powerful way to analyze data. However, not all quantities we care about have this qualitative feature of being clumped around a typical value.

Pareto's Principle (the 80–20 rule)

You may have heard of the so-called "**80–20 rule**" in business, with the tagline "*80% of sales come from 20% of customers*". This idea, however, did not come from sales and marketing. It originated from Vilfredo Pareto's study of Italian land ownership (circa 1890) [1].

Pareto observed that about 80% of the land in Italy was owned by about 20% of the population. It turns out that this simple observation indicates statistical properties that are *very different* than the Gaussian distributions we all know and love.

Namely, the "80–20 rule" is the consequence of a **Pareto distribution**. This is illustrated in the plot below.



Pareto distribution, where 20% of the population accounts for 80% of the volume. Image by author.

The **key difference** between a Gaussian and Pareto distribution is that a **Pareto does not have a "typical value"** that we can use to summarize the distribution efficiently.

In other words, while knowing the average weight of an Italian man (~175lbs) gives you a good idea of what to expect on your next trip to Rome, knowing the average population of an Italian city (~7,500) is **useless.**

Power Law Distributions

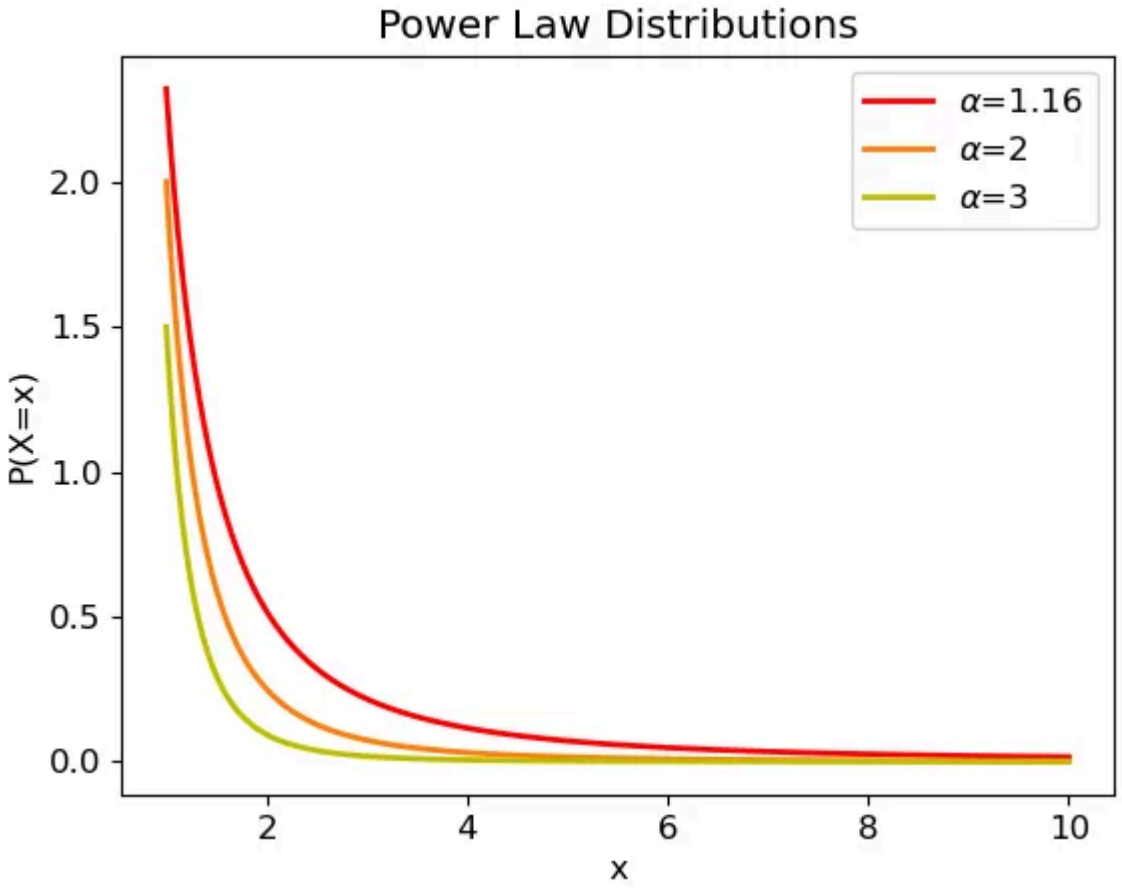
The Pareto distribution is part of a broader class of distributions called **Power Laws.** We can define a Power Law as follows [2].

$$PDF(x) = L(x)x^{-(\alpha+1)}$$

$$\text{Where,} \quad x > x_{min}$$

Definition of Power Law distribution class [3]. Image by author.

Where *PDF*() denotes the probability density function of a random variable, *X*. *x* is a particular value for *X*. *L(x)* is a slowly varying positive function with a domain of [*x_min,* ∞]. And *x_min* is the minimum value for which the power law holds (i.e. *PDF(x) = 0 for x<x_min*) [2]. And *α* is a number (typically between 2 and 3).



Example Power Law distributions with various α values. Note: α = 1.16 approximately implies the 80–20 rule. Image by author.

As we can see in the plots above, Power Laws are qualitatively very different from the Gaussian distribution. This forms a sort of **dichotomy between Gaussian-like and Pareto-like distributions.** Put another way, Gaussian and Power Law distributions provide conceptual anchors to qualitatively categorize things in the real world.

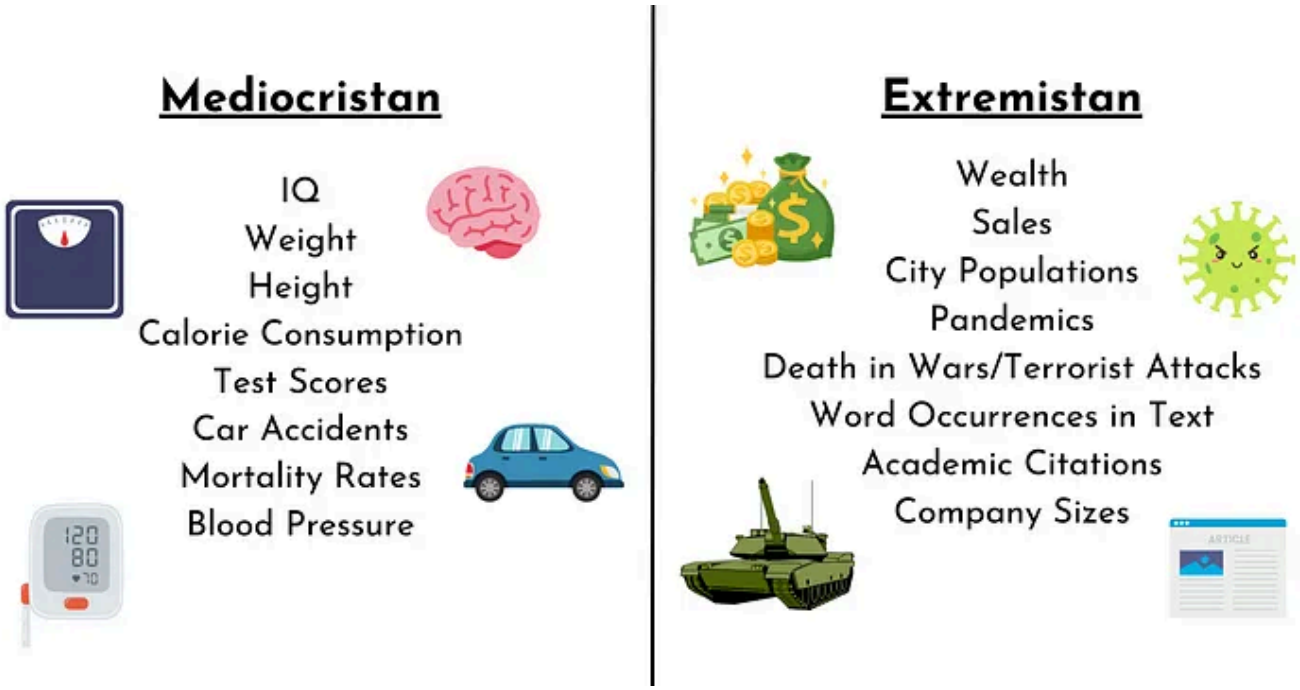
Mediocristan Vs Extremistan

Author Nassim Nicholas Taleb describes this dichotomy between Gaussian-like and Pareto-like things via **two categories** he calls "**Mediocristan**" and "**Extremistan**."

Mediocristan is the land of Gaussian-like things. A fundamental property of its citizens is **no single observation will significantly impact the aggregate statistics** [3]. For example, suppose you weigh every tourist at the Colosseum during your trip to Rome and compute the average weight. If you added the heaviest Italian on Earth, this average would be essentially unchanged (+0.5%).

On the other side of this conceptual landscape is **Extremistan**, where we see an opposite statistical property. Namely, in Extremistan, **a single observation can (and often will) drive the aggregate statistics.** Consider the same tourists at the Colosseum, but instead of measuring their weight, you ask each their net worth and compute the average. Unlike before, this average would change *dramatically* (+2500%) if we added the world's richest Italian, Giovanni Ferrero (the chocolate + hazelnut family), to the sample.

To get a better intuition for each of these categories, consider the examples listed in the image below.



Items from Mediocristan and Extremistan, respectively [3]. Image by author.

As you can see, the **Pareto-like inhabitants of Extremistan are not a small or trivial set.** In fact, many things we care about are **not** like the Gaussian curves we study in STAT 101.

While this may seem overly technical and didactic, there are **major limitations** in using our familiar statistical techniques and intuitions to analyze data generated from Extremistan and even (in some cases) significant risks.

3 Problems with STAT 101 Thinking

As we saw at the Roman Colosseum, data generated from Mediocristan (e.g. weight) have opposite properties from Extremistan (e.g wealth).

One of the biggest problems in using STAT 101 techniques to analyze Power Laws (i.e. data from Extremistan) is quantities like mean, standard deviation, variance, correlation, etc. all **have little practical significance.**

This all stems from a single core issue — **insufficient data.**

In statistics, we learn about the **Law of Large Numbers,** which says **that if we take N random samples,** the sample mean will approach the *true* mean as $N \to \infty$. This is true for ANY distribution (with finite mean): Gaussian, Power Law, Uniform, you name it.

However, it turns out that **this asymptotic behavior happens more *slowly* for some distributions than others** (e.g. slower for Power Laws than Gaussians). And, in practice, where we (necessarily) have finite datasets, this can cause problems. Here, I highlight 3 such problems.

Problem 1: The Mean is Meaningless (as well as many other metrics)

Whenever we want to compare two sets of values (*e.g. sales in April vs. May, traffic accidents in LA vs. NYC, patient outcomes in the control vs. treatment group*), we often compute a mean. This gives us an intuitive way to compress several values into a single representative number.

This works incredibly well for data that follow a nice Gaussian distribution because one can accurately estimate the mean in small sample sizes (N=~10). **However, this approach breaks down** when working with data following a Power Law distribution.

We can see this by comparing Gaussian and Power Law sample means as sample size increases, as shown in the plots below for N=100, N=1,000, and

N=10,000. Power Law and Gaussian sample means are plotted in orange and blue, respectively.

Sample mean convergence for 3 different sample sizes. Image by author.

As we can see, the Power Law sample means are more erratic (and biased) than the Gaussian. Even when the sample size is increased to N=100,000, the Power Law's accuracy is still much worse than what we see in the Gaussian for N=100. This is shown in the plot below.

Although the mean somewhat stabilizes at N=1,000,000, it is still significantly biased compared to the Gaussian. Image by author.

This erratic behavior is not only limited to the mean. It also applies to many commonly used statistical quantities. Similar convergence plots for the median, standard deviation, variance, min, max, 1st and 99th percentiles, kurtosis, and entropy are given below.

Other metric convergence plots at 3 sample sizes. From top to bottom: median, standard deviation, variance, min, max, 1st and 99th percentiles, kurtosis, and entropy. Image by author.

As we can see, **some metrics tend to be more stable than others.** For instance, the median, minimum, and percentiles hold up relatively well. Meanwhile, the standard deviation, variance, maximum, kurtosis, and entropy can't seem to settle on a single number.

Of this latter group, I want to point out the maximum because this quantity **might seem to converge in a small sample, but as N gets bigger, it can jump up by an order of magnitude** (as seen in the N=10,000 plot). This is especially dangerous because it can lead to a false sense of predictability and safety.

To tie this to the real world, if the underlying data were, say, deaths from a pandemic, the biggest pandemic in 100 years would be 10X smaller than the biggest one over 1,000 years.

For instance, the deadliest pandemic in the past 100 years was the Spanish flu (~50 million deaths) [4], so if deaths from a pandemic follow a Power Law distribution, then we can expect a pandemic claiming 500 million lives within the next 1,000 years (sorry for the dark example).

This highlights the key property of data from Extremistan, which is **rare events drive the aggregate statistics.**

However, this doesn't stop with the statistical metrics presented here. The gravity of rare events also impacts our ability to make predictions effectively.

Problem 2: Regression Doesn't Work

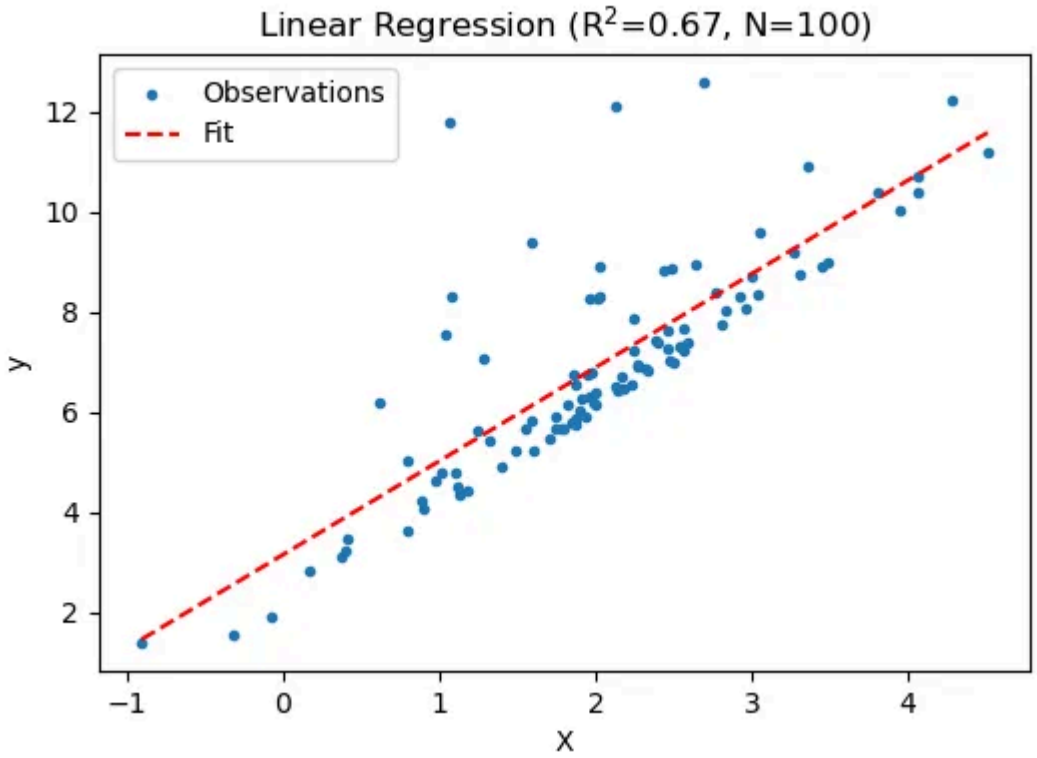
Regression boils down to making predictions based on past data. However, as we saw in Problem 1, when dealing with Power Laws we may not have sufficient data to accurately capture the *true* statistics.

This point is exacerbated when doing regression with variables following a Power Law distribution with $\alpha <= 2$. This is because an $\alpha <= 2$ **implies** the distribution has an **infinite variance**, which blows up a key assumption of popular regression methods (e.g. least squares regression).

However, when working with data in practice, one will never compute an infinite variance (the data is necessarily finite). This raises an issue similar to Problem 1: **the results may appear stable but do not hold up as you collect more data.**

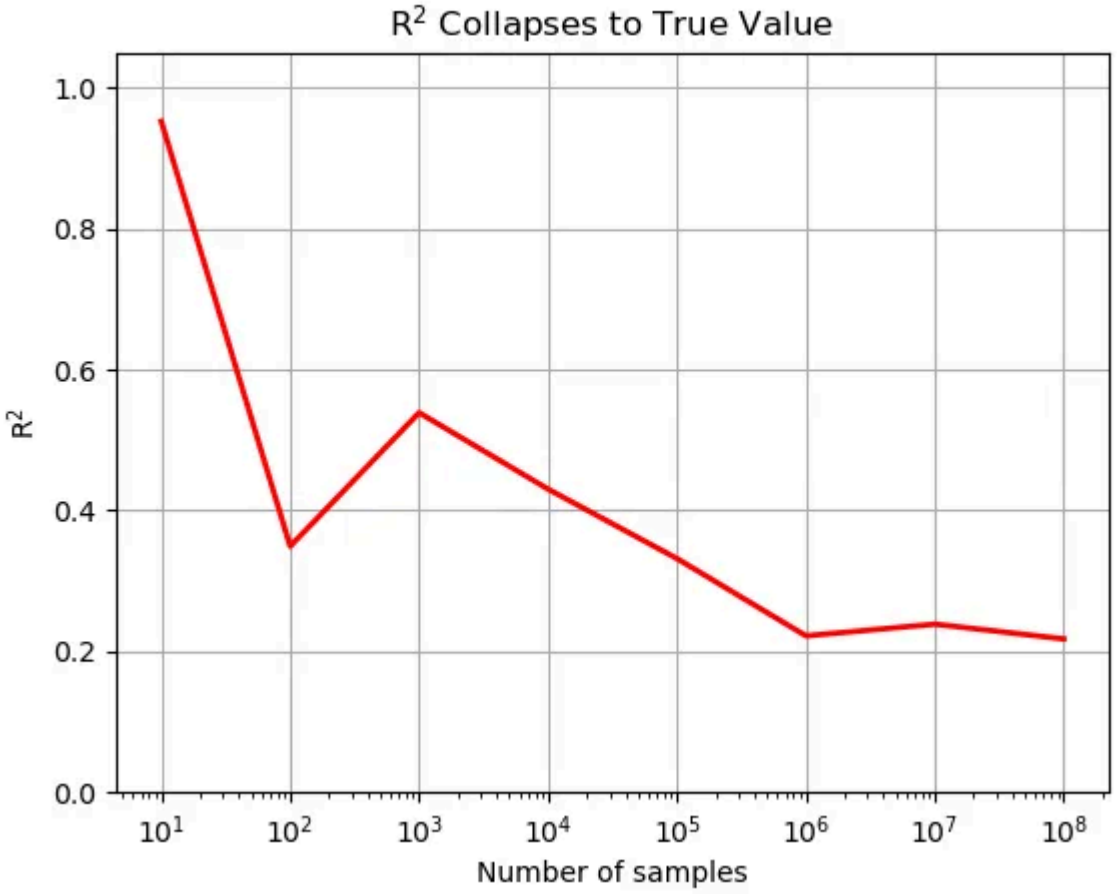
Put another way, your R^2 may look great when developing your model but quickly deteriorates as the sample size increases and approaches the actual value of $R^2=0$.

We can see this by way of an (artificial) example. Suppose we have two variables, X and Y, who are linearly related (i.e. Y = mX + b), where X is normally distributed with an additive noise term that follows a Power Law distribution. When we perform a regression in a small sample size (N=100), the fit performs deceivingly well.



Linear regression fit involving predictor with additive noise following a Power Law for a small sample size (N=100). Image by author.

However, as we collect more data (N=100,000,000), the R^2 correctly drops toward the actual value (i.e. R^2=0).



R² approaches actual value (i.e. R²=0) as sample size increases. Image by author.

Problem 3: Payoffs Diverge from Probabilities

At this point, you might think, "*Shaw.. what's the big deal? So what if my model can't predict a few rare events? It's right most of the time.*"

I agree with you. When working with data from Extremistan, it's easy to be right most of the time since most data are not in the tail. However, probabilities are only half the story when predicting outcomes and making decisions.

The other half of the story is payoffs. In other words, it's not just about *how often* you are right (wrong) but also *what happens* when you're right (wrong).

For instance, if offered a daily multivitamin that works great 99.9% of the time but kills you 0.1% of the time, you'd probably go with another brand (or eat better foods).

Relying solely on probabilities to make decisions is **especially detrimental when dealing with Power Laws and "80–20 rules"**. Consider the following business example.

Suppose we have a software company with 3 offerings: 1) free with ads, 2) premium, and 3) enterprise, where the distributions of customers and revenue for each offering are shown in the table below.

	Offer 1 Free with Ads	**Offer 2** Premium	**Offer 3** Enterprise
% of Customers	80%	16%	4%
% of Revenue	20%	16%	64%

Customer and Revenue distributions per offer. Image by author.

The company wants to roll out an update to speed up processing time by 50%. Being a cutting-edge, data-driven technology company, they surveyed active users and found that **95% of customers preferred the updated software**. With the data in hand, the company green lights the software update.

Six weeks later, however, the company is in disarray because revenue dropped 50%.

It turned out that after the update, 3 customers dropped the service because the update removed legacy data integrations that were essential to their use case. But these weren't just any customers. These were the company's **top 3 clients (~1%), making up about 50% of its revenue** (after all their custom upsells).

This is the kind of (fatal) mistake one can make when only focusing on probabilities (95% of customers loved the update). The moral of the story is when dealing with rare-event-driven data from Extremistan, **being wrong 1 time can cancel out being right 99 times (and then some).**

Code to generate plots 👇

Controversy In Extremistan

Power Laws, like Gaussians, are an idealized mathematical abstraction. The real world, however, is messy and rarely (if ever) will completely conform to our beautiful and precise constructions. This has raised some controversy about whether a particular distribution is *truly* a Power Law.

One point of debate has been whether wealth is a Power Law (as suggested by Pareto's work) or merely a log-normal distribution [5].

Some of the controversy might be explained by the observation that log-normal distributions behave like Gaussian for low sigma and like Power Law at high sigma [2].

However, to avoid controversy, we can depart (for now) from *whether some given data fits a Power Law or not* and focus instead on **fat tails.**

Fat-tailedness — measuring the space between Mediocristan and Extremistan

Fat Tails are a more general idea than Pareto and Power Law distributions. One way we can think about it is that "**fat-tailedness**" is the **degree to which rare events drive the aggregate statistics of a distribution**. From this point of view, fat-tailedness lives on a spectrum from not fat-tailed (i.e. a Gaussian) to very fat-tailed (i.e. Pareto 80–20).

This maps directly to the idea of Mediocristan vs Extremistan discussed earlier. The image below visualizes different distributions across this conceptual landscape [2].

Map of Mediocristan and Extremistan. Note: Since fat-tailedness lives on a spectrum, labeling a distribution as "Fat Tailed" or not is somewhat subjective. Image by author.

While there is no exact measure of fat-tailedness, there are many metrics and heuristics we can employ in practice to get a sense of where a given distribution sits on this map of Mediocristan and Extremistan. Here are a few approaches.

- **Power Law-iness**: use Power Law tail index i.e. α — the lower the alpha, the fatter the tails [2]
- **Non-gaussianity**: Kurtosis (breaks down for Power Law with $\alpha \leq 4$)
- **Variance** of Log-normal distribution
- **Taleb's κ** metric [6]

Takeaways

The central challenge with fat-tailed data is that one may not always have sufficient data to capture their underlying statistical properties accurately. This informs a few takeaways I will leave for the data practitioner.

- **Plot distributions** e.g. histograms, PDFs, and CDFs
- Ask yourself — **is this data from Mediocristan or Extremistan** (or somewhere between)**?**
- When building models, ask yourself — **what's the value of a correct prediction and the cost of an incorrect one?**
- If working with (very) fat-tailed data, don't ignore the rare events. Instead, **figure out how to use them** (e.g. can you do a special promotion for your top 1% of customers to drive more business?)

👉 **More on Power Laws & Fat Tails:** Power Law Fits | Quantifying Fat Tails

Detecting Power Laws in Real-world Data with Python

Breaking down a Maximum Likelihood-based approach with example code

towardsdatascience.com

Resources

Connect: My website | Book a call | Ask me anything

Socials: YouTube 📽️ | LinkedIn | Twitter

Support: Buy me a coffee ☕

The Data Entrepreneurs

A community for entrepreneurs in the data space. 👉 Join the Discord!

medium.com

[1] Pareto principle. (2023, October 30). In *Wikipedia*. https://en.wikipedia.org/wiki/Pareto_principle

[2] arXiv:2001.10488 [stat.OT]

[3] *Taleb, N.N. (2007). The Black Swan: the impact of the highly improbable. New York; Random House.*

[4] https://www.archives.gov/exhibits/influenza-epidemic/

[5] arXiv:0706.1062 [physics.data-an]

[6] Taleb, N. N. (2019). How much data do you need? An operational, pre-asymptotic metric for fat-tailedness. *International Journal of Forecasting*, *35*(2), 677–686. https://doi.org/10.1016/j.ijforecast.2018.10.003

Data Science | Statistics | Power Law | Pareto Principle | Getting Started

Written by Shaw Talebi

Edit profile

6.8K Followers · Writer for Towards Data Science

Data Scientist | PhD, Physics | Editor for The Data Entrepreneurs

More from Shaw Talebi and Towards Data Science

Shaw Talebi in Towards Data Science

QLoRA—How to Fine-Tune an LLM on a Single GPU

An introduction with Python example code (ft. Mistral-7b)

✦ · 16 min read · Feb 22, 2024

726 2

Patrick Brus in Towards Data Science

How to Write Clean Code in Python

Top takeaways from the book Clean Code

✦ · 21 min read · 6 days ago

885 11

Hamza Gharbi in Towards Data Science

Building a Chat App with LangChain, LLMs, and Streamlit f...

Build and deploy a chat application for complex database interaction with LangChai...

16 min read · Feb 9, 2024

Shaw Talebi in Towards Data Science

How to Build an AI Assistant with OpenAI + Python

Step-by-step guide on using the Assistants API & Fine-tuning

✦ · 13 min read · Feb 8, 2024

1K 9 478 4

See all from Shaw Talebi | See all from Towards Data Science

Recommended from Medium

Louis Chan in Towards Data Science

SHAP: Explain Any Machine Learning Model in Python

Your Comprehensive Guide to SHAP, TreeSHAP, and DeepSHAP

· 13 min read · Jan 11, 2023

805 3

Ignacio de Gregorio

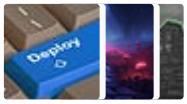
Google Has Finally Dethroned ChatGPT

They Finally Did It

· 10 min read · Feb 22, 2024

4.8K 90

Lists

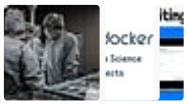


Predictive Modeling w/ Python

20 stories · 955 saves

Practical Guides to Machine Learning

10 stories · 1125 saves

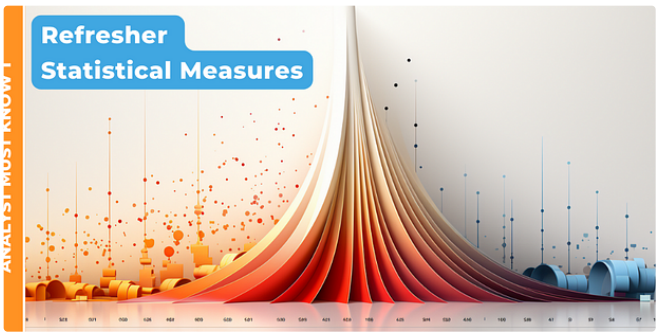


Coding & Development

11 stories · 473 saves

ChatGPT prompts

44 stories · 1187 saves



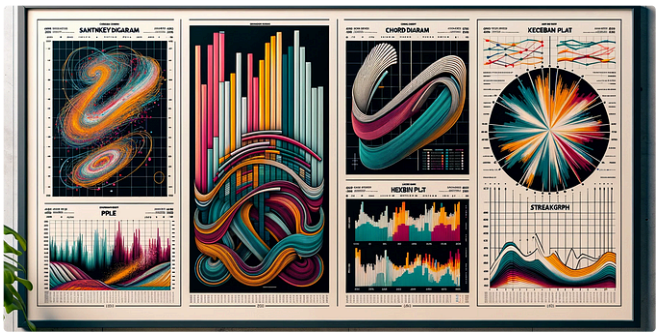
Prof. Frenzel

Statistical Measures Every Analyst Must Know—Part1

Measures of Central Tendency, Variability, Quartiles, Z-Scores, and as always:...

11 min read · Feb 4, 2024

283 1



Dr. Ashish Bamania in Level Up Coding

5 Extremely Useful Plots For Data Scientists That You Never Knew...

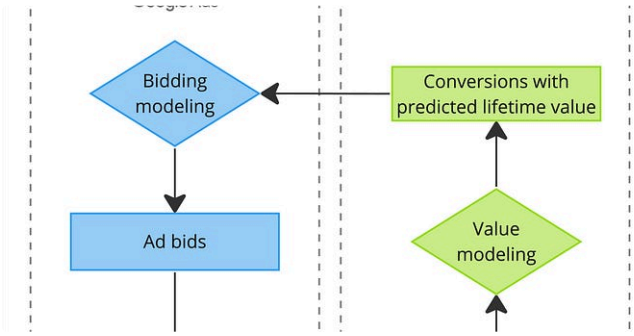
“5. Theme River”

· 6 min read · Jan 2, 2024

2.8K 19

Builescu Daniel in Python in Plain English

5 Python Coding Errors That Are Killing Your Speed (And How to Fi...



Nikolas Schriefer in HelloTech

How predicting customer lifetime value enables HelloFresh to...

Slow Python code? 5 easy fixes for instant speed-up.

· 6 min read · 6 days ago

611 1

Efficient Search Engine Advertising with tROAS optimization

7 min read · Feb 12, 2024

636 10

See more recommendations