The tail of a random variable

Alex Fu

2023-02-08

This note is rather hastily written, but at least in a jumbled mess you'll find a collection of related ideas.

Proceeding further along our quest to characterize and understand the distribution of a random variable, eventually, we would like to find a solution to the moment problem.

- We have seen that the distribution of a Borel- or Lebesgue-measurable real-valued random variable is uniquely characterized by its cumulative distribution function $F(x) = \mathbb{P}(X \leq x)$, or equivalently the complementary cdf (**ccdf**) $\bar{F}(x) = 1 - F(x) = \mathbb{P}(X > x)$, also known as the survival function or tail probability of X.

- We can decompose a distribution, a pushforward measure on $\mathbb{R}$, into a sum of three parts. First, a pure point measure containing the countably many atoms, which must be singleton points with mass $\mathbb{P}(X = x) > 0$ by the structure of the Borel σ-algebra. Then, an absolutely continuous measure, whose cdf is differentiable, with derivative the **pdf** $f(x) = \frac{\mathrm{d}}{\mathrm{d}x} F(x)$, representing its density relative to the canonical Lebesgue measure. Lastly, a singularly continuous measure, guaranteed by the Lebesgue–Radon–Nikodym decomposition, supported on a set of zero length.

  In some sense, a distribution is like a jar of chunky peanut butter spread over the bread of the real line, with jelly sticking "orthogonally" out of the bread.

- We defined the moment-generating function (mgf) $M_X(t) = \mathbb{E}(e^{tX})$ and characteristic function (ch.f), or Fourier transform $\varphi_X(t) = \mathbb{E}(e^{itX})$, the Fourier dual of the pdf. We considered issues of existence (around a neighborhood of 0), as well as various forms of regularity: continuity, uniform continuity, k-times continuous differentiability, integrability, etc., and connected these to the moments.

- We considered the moments $\mathbb{E}(X^k)$ as another possible way to characterize the distribution, and constructed the L^p spaces with the norm $\|X\|_p = \mathbb{E}(X^p)^{1/p}$. We saw the monotonicity of moments and linear inclusion of L^p spaces, and various inequalities on L^p and L^q.

- We also considered the class of $\mathbb{E}(g(X))$ for g bounded and continuous as a means to determine a distribution, which appears in the definition of weak convergence, the portmanteau lemma, and the definition of conditional expectation, where a random variable is $\sigma(Y)$-measurable iff it is a measurable function of Y.

Much of the analysis we have done so far is tied in one way or another to the **tail** of a distribution, $\mathbb{P}(|X| \geq t)$. Not only does the tail fully characterize the distributions of nonnegative or symmetric random variables, it provides very specific, useful quantitative information, combining asymptotic and probabilistic considerations.

(We will mostly be working with $|X|$ and $\mathbb{E}(|X|^p)$, so taking $X \mapsto |X|$, we will assume without loss of generality that X is nonnegative. One can envision this as folding the distribution in half along 0, and taking the maximum or heavier tail between X^+ and $-X^-$. Or, alternatively, one pretends it's a symmetric distribution, and correspondingly doubles the tail probability of a single branch.)

The tail gives a measure of **concentration**, e.g. through Markov's, Chebyshev's, Cantelli's, or Chernoff's inequalities, fully revealing the "distribution" — location *and* relative quantity — of probability mass as the threshold t sweeps through $\mathbb{R}^+$, or along a particular asymptotic trajectory.

The tail's heaviness or lightness, more specifically its asymptotic class as $t \to \infty$, determines the existence of moments. The zeroth moment lets the distribution sit; the first moment balances the distribution on a seesaw fulcrum; the second moment rotates the distribution along a vertical axis; and higher moments prioritize non-exceptional smallness in magnitude and punish large deviations, as if "vibrating" the distribution in further dimensions and assessing its "impact." The more probabilistically spread out the distribution, the larger higher-order moments are.

For a very basic result,

X is finite almost surely iff $\mathbb{P}(X > t) \in o(1)$.

By monotone continuity, $\lim_{t\to\infty} \mathbb{P}(X > t)/1 = \mathbb{P}(X = \infty)$, so $\mathbb{P}(X > t) \in o(1)$ iff $\mathbb{P}(X = \infty) = 0$; otherwise, $\mathbb{P}(t) \in \Theta(1)$, with $\mathbb{P}(X = \infty) = c > 0$.

We might imagine this result as "characterizing" the finiteness of the zeroth moment $\mathbb{E}(X^0)$, usually the probability mass $1 = \int_{-\infty}^{\infty} f(t)\,\mathrm{d}t$, but by taking the unusual convention of $\infty^0 = \infty$, we have a strange interpretation.

And, of course, X is bounded almost surely if there exists $t < \infty$ such that $\mathbb{P}(X > t) = 0$.

(An interesting unrelated tidbit that this reminds me of: mean minimizes the ℓ^2 norm, median the ℓ^1 norm, and *mode* the ℓ^0 norm, which simply counts or indicates whether a given sample is nonzero. So the usual statistics have some justification in the form of $\min|X - \hat{X}|$. Here, X^0 detects whether X is finite.)

Let's move on to a stronger condition than being finite a.s.: having first moment $\mathbb{E}\,|X| < \infty$, or being *integrable*. From our knowledge of p-series and integrals of x^p, we know that the divergent harmonic series $\sum_{n=1}^{\infty} n^{-1} = \infty$ and the Riemann integral $\int_1^{\infty} x^{-1}\,\mathrm{d}x = \ln\infty = \infty$, which are really at most 1 away from each other by left and right Riemann approximation, form a "boundary" separating convergence from divergence. A finer boundary is given by $(n \log n)^{-1}$, as we noted in *2023-01-18*.

Using the intuition of expectation as $X \cdot \mathbb{P}$, an accumulation of values weighted by probabilities, like sums or integrals, we might surmise $\sum_{n=1}^{\infty} n \cdot n^{-(2+\varepsilon)} < \infty$ or $\int_1^{\infty} x \cdot x^{-(2+\varepsilon)}\,\mathrm{d}x < \infty$ means we want the pmf or pdf in the order of $o(x^{-2})$, which makes the tail $\mathbb{P}(X > t) \prec \sum_{n>t} n^{-2} \approx \int_t^{\infty} x^{-2}\,\mathrm{d}x \sim t^{-1}$.

Is $\mathbb{P}(X > t) \in o(t^{-1})$ enough? Well, $t\,\mathbb{P}(X > t) \to 0$ does not imply integrability, though the converse is true by Markov's inequality: $t\,\mathbb{P}(X > t) \leq \mathbb{E}(X\mathbb{1}_{X>t}) \leq \mathbb{E}(X) < \infty$, where the upper bound tends to 0 by DCT. Markov's is only a one-sided bound, after all.

For a counterexample, consider discrete X that takes on each $n \geq 1$ with probability $n^{-2}(\log n)^{-1}$, normalized by the appropriate constant $C \leq \frac{\pi^2}{6}$. The expectation clearly diverges, but the tail $\mathbb{P}(X > n) \in o(n^{-1})$.

But, the question remains: is $\mathbb{P}(X > t) \in o((t \log t)^{-1})$ enough? If we forced the tail to be approximately summable or Riemann integrable, e.g. $t^{-1}(\log t)^{-(1+\varepsilon)}$, is this necessary and sufficient for $\mathbb{E}(X) < \infty$? To answer this question, we will first derive the tail-sum formula for discrete and continuous random variables, then the tail-sum approximation for the first moment in general.

For $\mathbb{N}$-valued X, $\mathbb{E}(X) = \sum_{n=1}^{\infty} \mathbb{P}(X \geq n) = \sum_{n=0}^{\infty} \mathbb{P}(X > n)$.

Consider the triangular array of values

$$\begin{array}{llll}
\mathbb{P}(X=1) & & & \\
\mathbb{P}(X=2) & \mathbb{P}(X=2) & & \\
\mathbb{P}(X=3) & \mathbb{P}(X=3) & \mathbb{P}(X=3) & \\
\mathbb{P}(X=4) & \mathbb{P}(X=4) & \mathbb{P}(X=4) & \mathbb{P}(X=4) \\
\vdots & & &
\end{array}$$

This represents all the values we wish to sum over. If we first sum along each row, then along the collapsed column, we find the usual definition of discrete expectation $\mathbb{E}(X) = \sum_{n=1}^{\infty} n\,\mathbb{P}(X = n)$. However, if we first sum along each column, then down the diagonal, we find the **tail sum** $\sum_{n=1}^{\infty} \mathbb{P}(X \geq 1)$.

This different row-column order is really interchanging the order of nested double summations, $\sum_{n=1}^{\infty} \sum_{k=1}^{n} \mathbb{P}(X = n) = \sum_{k=1}^{\infty} \sum_{n>k} \mathbb{P}(X = n)$, which is justified by Tonelli's theorem, where all values above are nonnegative. Thus, we can thus apply the tail-sum formula to $\mathbb{E}\,|X|$, or $\mathbb{E}\,X = \mathbb{E}\,X^+ - \mathbb{E}\,X^-$, but not directly to possibly negative X, where the issues of absolute and conditional convergence rear their heads.

The idea of exchanging the order of integration comes up again in the continuous. Consider the indicator function $\mathbb{1}_{x>t}$. For fixed x, it captures the region of t with ceiling x, t varying in $[0, x]$. But for fixed t, it indicates the semi-infinite ray (t, ∞).

For continuous X, $\mathbb{E}(X) = \int_0^{\infty} \mathbb{P}(X > t)\,\mathrm{d}t = \int_0^{\infty} \mathbb{P}(X \geq t)\,\mathrm{d}t$.

This is just the continuous analogue of the discrete triangular array. Consider the usual definition of continuous expectation, which we can write suggestively as an iterated integral:

$$\mathbb{E}(X) = \int_0^{\infty} xf(x)\;\mathrm{d}x = \int_0^{\infty} \int_0^{x} f(x)\;\mathrm{d}t\;\mathrm{d}x = \int_0^{\infty} \int_0^{\infty} f(x)\mathbb{1}_{x>t}\;\mathrm{d}t\;\mathrm{d}x$$

The "number" or weight of $f(x)$ is x, the length of the region $(0, x]$, just as the weight for $\mathbb{P}(X = n)$ was n, the count of the values in $(0, n]$.

Again, by the analogous Fubini's theorem, we wish to interchange the order of the two integrals:

$$\int_0^{\infty} \int_0^{\infty} f(x)\mathbb{1}_{x>t}\;\mathrm{d}t\;\mathrm{d}x = \int_0^{\infty} \int_0^{\infty} f(x)\mathbb{1}_{x>t}\;\mathrm{d}x\;\mathrm{d}t = \int_0^{\infty} \int_t^{\infty} f(x)\;\mathrm{d}x\;\mathrm{d}t = \int_0^{\infty} \mathbb{P}(X > t)\;\mathrm{d}t.$$

Alternatively, the tail-sum formula is also an integration by parts representation, from continuous calculus or the calculus of finite differences (e.g. Abel's summation by parts). $u\,\mathrm{d}v = \mathrm{d}(uv) - v\,\mathrm{d}u$ expresses a tradeoff: one power is lowered to raise another power; one term x differentiated to integrate the other term f.

We have to be a bit careful, however: the naïve approach of simply taking

$$\int xf\;\mathrm{d}x = \int x\;\mathrm{d}F = -\int x\;\mathrm{d}\bar{F} = \int x\bar{F}\;\mathrm{d}x - \int 1\bar{F}\;\mathrm{d}x = -\int \bar{F}\;\mathrm{d}x?$$

will not work without more careful consideration of the bounds of integration involved.

The tail, the asymptotic rate of decay of the probability of deviation, is closely tied to the existence of moments, and its associated concepts such as the regularity of the ch.f and L^p inequalities.

I may have forgotten to mention, but the tail is called the tail because it is graphically the "tail" of the pmf or pdf for a generic distribution, which is thought of as vaguely bell curve in shape.

We will continue for a hopefully less chaotic derivation for an approximation for $\mathbb{E}\, X^p$, $p \geq 1$, which by integration by parts we might expect to be $\sim \int p x^{p-1} \bar{F} \, \mathrm{d}x$.

■