Note 07. Concentration inequalities

Alex Fu

Fall 2022

Another intuitive feature of distributions is *concentration*, the tendency for most probability mass to be around the center, or the opposite *anti-concentration*. **Concentration inequalities** are (typically upper) bounds on the probability that a random variable deviates from some value, usually its mean, often formulated in terms of its moments.

For a related idea, consider **confidence intervals** for the normal distribution $\mathcal{N}(\mu, \sigma^2)$. Given some fixed probability or *confidence level* $C \in [0, 1]$, we wish to find the *critical value* z^* so that we are C confident of being at most z^* standard deviations from the mean:

$$\mathbb{P}(X \in [\mu - z^* \cdot \sigma, \mu + z^* \cdot \sigma]) = C.$$

If Φ^{-1} is the standard normal cdf, we can find z^* graphically as

$$z^* = \frac{1}{\sigma}\left[\mu - \Phi^{-1}\left(\frac{1-C}{2}\right)\right].$$

With concentration inequalities, our task is almost the reverse: given some deviation, we wish to bound the probability of deviation, now without knowing anything specific about the distribution.

1 Markov's inequality

The simplest concentration inequality is Markov's inequality, which only uses information about the first moment. Other inequalities that incorporate information about higher moments often, but not always, provide tighter bounds. However, the weaker assumptions of Markov's inequality make it powerful in being more generalizable.

Proposition 1 (Markov's inequality)**.**

Let X be an almost surely nonnegative random variable. Then for any $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Proof. We observe the following indicator inequality, then apply the monotonicity of expectation:

$$\mathbb{1}\{X \geq a\} \leq \frac{X}{a}.$$

Alternatively, we can examine the definition of expectation:

$$\mathbb{E}(X) = \mathbb{E}(X \cdot \mathbb{1}_{X<a}) + \mathbb{E}(X \cdot \mathbb{1}_{X \geq a}) \geq 0 \cdot \mathbb{P}(X < a) + a \cdot \mathbb{P}(X \geq a).$$

□

We are not limited to nonnegative random variables. Every random variable X can be split into positive and negative parts, $X^+ := \max(X, 0)$ and $X^- := -\min(X, 0)$. Then $X^+, X^- \geq 0$, $X = X^+ - X^-$, and $|X| = X^+ + X^-$. Thus, for any X,

$$\mathbb{P}(X \geq a) \leq \mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}(|X|)}{a}.$$

In the special case that X is symmetric about 0, the usual Proposition 1 still applies.

In fact, more can be said about Markov's inequality. If $\phi\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is *monotonically increasing*, then $X \geq a$ implies $\phi(X) \geq \phi(a)$. If ϕ is also *invertible*, then the converse is true as well, so $X \geq a$ and $\phi(X) \geq \phi(a)$ are equivalent events.

Proposition 2 (Generalized Markov's inequality)**.**

Let X be as in Proposition 1, and let $\phi\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a strictly monotonically increasing function. Then for any $a > 0$,

$$\mathbb{P}(X \geq a) = \mathbb{P}(\phi(X) \geq \phi(a)) \leq \frac{\mathbb{E}(\phi(X))}{\phi(a)}.$$

In particular, $\phi(X) = |X|^n$ for $n \geq 1$ is strictly monotonically increasing.

2 Chebyshev's inequality

Proposition 3 (Chebyshev's inequality)**.**

Let X be a random variable with finite second moment. Then for any $a > 0$,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathsf{var}(X)}{a^2}.$$

Chebyshev's inequality follows by applying the generalized Markov's inequality to $|X - \mathbb{E}(X)|$, which gives a better result than applying it to $|X|$, as $\mathsf{var}(X) \leq \mathbb{E}(X^2)$. Equivalently,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq z \cdot \sigma) \leq \frac{1}{z^2}.$$

Find some examples where Chebyshev's inequality is strict. Does it necessarily give a better bound than Markov's inequality?

Proposition 4 (Cantelli's inequality, or one-sided Chebyshev's inequality*).

Let X have finite variance σ^2. Then for any $a > 0$,

$$\mathbb{P}(X - \mathbb{E}(X) \geq a) \leq \frac{\sigma^2}{\sigma^2 + a^2}.$$

Proof. Consider the monotonically increasing function $\phi(X) = (X - \lambda)^2$ for $\lambda \geq 0$. Noting that $X' = X - \mathbb{E}(X)$ also has variance σ^2, we apply the generalized Markov's inequality:

$$\mathbb{P}(X - \mathbb{E}(X) \geq a) = \mathbb{P}(\phi(X') \geq \phi(a)) \leq \frac{\mathbb{E}((X' + \lambda)^2)}{(a + \lambda)^2} = \frac{\sigma^2 + \lambda^2}{(a + \lambda)^2}.$$

A common technique: as the probability is independent of λ, and the upper bound holds for all λ, the inequality must hold for the least upper bound as well. We can find the minimum or infimum by differentiation with respect to λ:

$$\mathbb{P}(X - \mathbb{E}(X) \geq a) \leq \inf_{\lambda \geq 0} \frac{\sigma^2 + \lambda^2}{(a + \lambda)^2} = \left.\frac{\sigma^2 + \lambda^2}{(a + \lambda)^2}\right|_{\lambda = \frac{\sigma^2}{a}} = \frac{\sigma^2}{\sigma^2 + a^2}.$$

□

Another optional application of concentration inequalities can be used to show that a nonnegative random variable has a nonzero probability of being positive.

Proposition 5 (First moment method*).

Let X be an $\mathbb{N}$-valued random variable. Then

$$\mathbb{P}(X > 0) \leq \mathbb{E}(X).$$

The proof follows from Markov's inequality and $\mathbb{P}(X > 0) = \mathbb{P}(X \geq 1)$.

Proposition 6 (Second moment method*).

Let X be nonnegative with finite second moment. Then

$$\mathbb{P}(X > 0) \geq \frac{\mathbb{E}(X)^2}{\mathbb{E}(X^2)} = \frac{\mu^2}{\mu^2 + \sigma^2}.$$

The proof uses the Cauchy-Schwarz inequality for expectation:

$$\mathbb{E}(X)^2 = \mathbb{E}(X \cdot \mathbb{1}_{X>0})^2 \leq \mathbb{E}((X \cdot \mathbb{1}_{X>0})^2) = \mathbb{E}(X^2) \cdot \mathbb{P}(X > 0).$$

3 Chernoff bound

Inequalities incorporating second moments are often better than those only using first moments. What if we somehow incorporated information about every moment? Noting that the function $\phi(X) = e^{tX}$ for parameter $t > 0$ is strictly monotonically increasing,

Proposition 7 (Chernoff bound).

Let X be any random variable, and let $a \in R$. Then for any $t > 0$,

$$\mathbb{P}(X \geq a) = \mathbb{P}\big(e^{tX} \geq e^{ta}\big) \leq \frac{\mathbb{E}(e^{tX})}{e^{ta}}.$$

The proof follows from the generalized Markov's inequality. See this remark for the reason we can optimize the Chernoff bound by taking the infimum over all parameters $t > 0$ to the moment-generating function $\mathbb{E}(e^{tX})$. In general, we can even optimize over the class of strictly monotonically increasing functions $\phi\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$.

There are many other variations on the Chernoff bound, just as there are a great deal of concentration inequalities out of our scope that we nonetheless invite the interested reader to explore.

4 Hoeffding bound*

The laws of large numbers (presented in the following note) describe an important class of distributions which have an eventual concentration about their mean: the sum

$$S_n = X_1 + \cdots + X_n$$

of random variables $X_1, \ldots, X_n$, often with some conditions of independence and regularity, or alternatively the average $\bar{X}_n$. With concentration inequalities, we can more closely quantify the rate of concentration by some sharp upper bounds on the probability of deviation from the mean.

Proposition 8 (kth moment bounds).

Let $\lambda > 0$. By Markov's inequality,

$$\mathbb{P}(|S_n| \geq \lambda) \leq \frac{1}{\lambda} \sum_{i=1}^{n} \mathbb{E}(|X_i|).$$

If the X_i are pairwise independent, then by Chebyshev's inequality,

$$\mathbb{P}(|S_n| \geq \lambda) \leq \frac{1}{\lambda^2} \sum_{i=1}^{n} \mathrm{var}(X_i).$$

More generally, if the X_i are k-wise independent, then

$$\mathbb{P}(|S_n| \geq \lambda\sqrt{n}) \leq 2\lambda^{-k}\left(\frac{ek}{2}\right)^{k/2}$$

by Stirling's formula, where the kth moment of $|S_n|$ is

$$\mathbb{E}(|S_n|^k) = \sum_{1\leq i_1\leq\cdots\leq i_k\leq n} \mathbb{E}(X_{i_1}\cdots X_{i_k}).$$

As we have already seen, the moment-generating function $\mathbb{E}(e^{tS_n})$ may provide a better bound than any individual kth moment $\mathbb{E}(|S_n|^k)$.

Proposition 9 (Hoeffding lemma).

Let X take values in $[a,b]$ almost surely. Then for any $t>0$,

$$\mathbb{E}(e^{tX}) \leq e^{t\,\mathbb{E}(X)}\big[1+O\big(t^2\,\mathsf{var}(X)\cdot e^{O(t(b-a))}\big)\big].$$

In particular, the inequality remains true for $e^{t\,\mathbb{E}(X)}\cdot e^{O(t^2\,\mathsf{var}(X))}$.

Proof. Without loss of generality, we can translate and scale X so that $\mathbb{E}(X)=0$ and $b-a=1$. Then $X\in O(1)$, and by Taylor expansion,

$$\begin{aligned}\mathbb{E}(e^{tX}) &= \mathbb{E}\big(1+tX+O\big(t^2X^2\cdot e^{O(tX)}\big)\big)\\ &= 1+O\big(t^2\,\mathsf{var}(X)\cdot e^{O(t)}\big).\end{aligned}$$

We are done after reversing any transformations of X. For the more specific inequality, we note that $\mathsf{var}(X)\leq(b-a)^2$, so we can substitute $\mathsf{var}(X)$ by $(b-a)^2$. Then

$$O\big(1+t^2(b-a)^2e^{O(t(b-a))}\big)\subseteq O\Big(e^{t^2(b-a)^2}\Big)=e^{O(t^2\,\mathsf{var}(X))}$$

as the function $1+y^2e^y$ is dominated by e^{y^2}, where we take $y=t(b-a)$. □

Proposition 10 (Sharpened Chernoff bound).

Let $X_1,\ldots,X_n$ be independent random variables that take values in $[-K,K]$ almost surely, let $\mu=\mathbb{E}(S_n)$, and let $\sigma^2=\mathsf{var}(S_n)$. Then for any $\lambda>0$, there exist $C,c>0$ so that

$$\mathbb{P}(|S_n-\mu|\geq\lambda\sigma)\leq C\cdot\max\Big(e^{-c\lambda^2},e^{-c\lambda\sigma/K}\Big).$$

Proof. Let $\mu_i=\mathbb{E}(X_i)$, and let $\sigma_i^2=\mathsf{var}(X_i)$. Without loss of generality, we assume that $\mu_i\equiv 0$

and $K = 1$, so that $|X| \leq 1$. Then, by independence and the Hoeffding lemma,

$$\mathbb{E}(e^{tS_n}) = \prod_{i=1}^{n} \mathbb{E}(e^{tX_i}) \leq \prod_{i=1}^{n} e^{O(t^2\sigma_i^2)} = e^{\sum_{i=1}^{n} O(t^2\sigma_i^2)} = e^{O(t^2\sigma^2)}.$$

By the generalized Markov's inequality, optimized over the parameter $t \in [0, 1]$,

$$\mathbb{P}(S_n \geq \lambda\sigma) \leq \frac{\mathbb{E}(e^{tS_n})}{e^{t\lambda\sigma}} \leq e^{O(t^2\sigma^2) - t\lambda\sigma},$$

from which we obtain the desired inequality. □

Now let the parameter t take values outside of $[0, 1]$. Show that $e^{-c\lambda\sigma/K}$ can be replaced with

$$\left(\frac{\lambda K}{\sigma}\right)^{-c\lambda\sigma/K},$$

which is more optimal when $\lambda K \gg \sigma$.

Finally, from Proposition 10, we obtain the following concentration inequality.

Proposition 11 (Hoeffding bound).

Let $X_1, \ldots, X_n$ be independent random variables such that X_i takes values in $[a_i, b_i]$ almost surely, and let $\sigma^2 = \sum_{i=1}^{n} |b_i - a_i|^2$. Then there exist $C, c > 0$ such that

$$\mathbb{P}(|S_n - \mathbb{E}(S_n)| \geq \lambda\sigma) \leq Ce^{-c\lambda^2}$$

■