Probability Theory & Statistics

Inequalities & Limits

Inequalities

Markov. For any random variable X and constant $a > 0$,

$$P\left(|X| \geq a\right) \leq \frac{E\left(|X|\right)}{a} \tag{1}$$

For an intuitive interpretation, let X be the income of a randomly selected person from the population. Taking $a = 2E(X)$

$$P\left(|X| \geq 2E(X)\right) \leq \frac{1}{2} \tag{2}$$

which says that the probability an observed income is greater than twice the average must be less than $\frac{1}{2}$, i.e., it is impossible for more than half the population to make at least twice the average income.

For concreteness, imagine a sample, 0, 0, 2, 2, with average income of 1. Half of the population have exactly twice the average income. For the average to remain 1, either the income must be lowered and shared between the other people, e.g., 1, 1, 1, 1, otherwise the income is increased in a single person, e.g., 0, 0, 0, 4, decreasing the probability that X exceeds the average. Any other combination must result in a change in the average. Similarly, you can't have more than a third of the population earn three-times the average salary, i.e., $P\left(|X| \geq 3E\left(|X|\right)\right) \leq 1/3$.

Markov's inequality is quite crude since $E(|X|)/a$ can be any value, e.g., 1 or infinity (think Cauchy), requiring no assumptions about X. Surprisingly, from Markov's inequality we can derive two more useful bounds with no additional work.

Chebyshev. Let X have mean, μ, and variance, σ^2. Then for $a > 0$

$$P\left(|X - \mu| \geq a\right) \leq \frac{\sigma^2}{a^2} \tag{3}$$

Exercise. Prove Chebyshev's inequality using Markov's inequality.

For $a = c\sigma$, $c > 0$, we have the following equivalent inequality.

$$P\left(|X - \mu| \geq c\sigma\right) \leq \frac{1}{c^2} \tag{4}$$

This gives an upper bound on the probability of a random variable being more than c standard deviations away from its mean, e.g., there can't be more than a 25% chance of being 2 or more standard deviations away from the mean.

Weak Law of Large Numbers (LLN)

Weak law of large numbers. For all $\varepsilon > 0$,

$$P\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) \to 0, \tag{5}$$

2

as $n \rightarrow \infty$, where $n\overline{X}_n = X_1 + X_2 + \cdots$ is the *sample mean* of n IID random variables.

Exercise. Prove the weak LLN using Chebyshev's inequality.

The law of large numbers is essential for simulations, statistics, and science. Consider generating "data" from a large number of independent replications of an experiment, performed either by computer simulation or in the real world. Every time we use the average value in the replications of some quantity to approximate its theoretical average, we are implicitly appealing to the LLN.

Coin Tosses

Let X_1, X_2, ... be IID $Bern(1/2)$. Interpreting the X_j as indicators of heads in a string of fair coin tosses, where the sample mean is the proportion of heads after n tosses. The weak LLN says that for any $\varepsilon > 0$, the probability of the sample mean being more than ε away from $1/2$ can be made as small as we like by letting n grow.

In the supplementary code we use R to simulate sequences of coin tosses and compute their average as a function of n, the number of tosses.

Exercise. Does LLN contradict the fact that the coin is memoryless?

Central Limit Theorem

As in the previous example, let X_1, X_2, ... be IID with mean μ and variance σ^2. The LLN says that $\overline{X}_n$ converges to μ as $n \rightarrow \infty$, but what is the distribution of $\overline{X}_n$? The *central limit theorem* (CLT) provides the answer to this question.

Central limit theorem. As $n \rightarrow \infty$,

$$\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \rightarrow N(0, 1) \tag{6}$$

In words, the CDF (cumulative probability distribution) of the l.h.s. converges to the standard normal distribution.

Proof. Assume the *moment generating function* $M(t) = E(e^{tX_j})$ exists. Let $M(t) = E(e^{tX_j})$, and assume without loss of generality that $\mu = 0$, and $\sigma^2 = 1$. Then,

$$\begin{aligned} M(0) &= 1, \\ M'(0) &= E(X_j) = \mu = 0, \\ M''(0) &= E\left(X_j^2\right) = \sigma^2 = 1. \end{aligned} \tag{7}$$

We want to show that the moment generating function (MGF) of $\sqrt{n}\overline{X}_n = (X_1 + X_2 \ldots)/\sqrt{n}$ converges to the MGF of the Normal which is $e^{t^2/2}$.

By the property $M(f(X_1 + X_2)) = M(f(X_1) + f(X_2))$

$$E\left(e^{t(X_1+X_2\dots)/\sqrt{n}}\right) = E\left(e^{\frac{tX_1}{\sqrt{n}}}\right) E\left(e^{\frac{tX_2}{\sqrt{n}}}\right)\dots$$

$$= \left(M\left(\frac{t}{\sqrt{n}}\right)\right)^n \tag{8}$$

Letting $n \to \infty$, we get the indeterminate form 1^∞, so instead we take the logarithm, $n \log M\left(\frac{t}{\sqrt{n}}\right)$ and then exponentiate at the end

$$
\begin{aligned}
\lim_{n\to\infty} n \log M\left(\frac{t}{\sqrt{n}}\right) &= \lim_{y\to 0} \frac{\log M(yt)}{y^2} && \text{for } y = 1/\sqrt{n} \\
&= \frac{t}{2} \lim_{y\to 0} \frac{tM'(yt)}{2yM(yt)} && \text{by L'Hopital's rule} \\
&= \frac{t}{2} \lim_{y\to 0} \frac{M'(yt)}{y} && \text{since } M(yt) \to 1 \\
&= \frac{t^2}{2} \lim_{y\to 0} M''(yt) && \text{by L'Hopital's rule}
\end{aligned}
\tag{9}
$$

which gives $t^2/2$, hence the MGF approaches the $N(0,1)$ MGF.

Take a moment to understand the generality of this result. The distribution of the individual X_j can be anything in the world, as long as the mean and variance are finite. We could have a discrete distribution like the Binomial, a bounded distribution like the Beta, a skewed distribution like the Log-Normal, or a distribution with multiple peaks and valleys. No matter what, the act of averaging will cause Normality to emerge.

This does not mean that the distribution of the X_j is irrelevant, however. If the distribution is highly skewed or multimodal distribution, we may need n to be very large before the Normal approximation becomes accurate.

Coin Tosses (Revisited)

In the previous example we used the LLN to conclude that $\overline{X}_n$, the running proportion of heads, converges to $\frac{1}{2}$ for a large number of tosses. Using the CLT, we can now say more, i.e., that $E\left(\overline{X}_n\right) = 1/2$ and $Var\left(\overline{X}_n\right) = 1/(4n)$, so for large n

$$\overline{X}_n \sim N\left(\frac{1}{2}, \frac{1}{4n}\right) \tag{10}$$

This information allows us to quantify the sample mean's deviation from the population mean.

The CLT also says that the sample mean is approximately Normal with mean $n\mu$ and variance $n\sigma^2$.

Exercise. Show that this is true by simulation in R.

Discussion

Chebyshev's inequality says that for the random variable X with mean μ and variance σ^2, for $a > 0$,

$$P\left(|X - \mu| \geq a\right) \leq \frac{\sigma^2}{a^2} \tag{11}$$

From Markov's inequality we have

$$P\left(|Z| \geq a\right) \leq \frac{E\left(|Z|\right)}{a} \tag{12}$$

For $Y = X - \mu$ and $Z = Y^2$, therefore,

$$P\left(|Y| \geq a\right) = P\left(Y^2 \geq a^2\right) = P\left(Z \geq a^2\right) = \frac{E(Z)}{a^2} = \frac{E(X - \mu)^2}{a^2} = \frac{\sigma^2}{a^2} \tag{13}$$

The Weak Law of Large Numbers (WLLN) says that for all $\varepsilon > 0$, $P\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) \to 0$ as $n \to \infty$. *Proof.* Fix ε to some positive number. By Chebyshev's inequality

$$P\left(|Y - \mu| \geq c\sigma_Y\right) \leq \frac{1}{c^2} \tag{14}$$

with $Y = \overline{X}_n$ and $c = \frac{\varepsilon}{\sigma}$ we obtain

$$P\left(\left|\overline{X}_n - \mu\right| \geq \varepsilon\right) \leq \frac{1}{\left(\frac{\varepsilon}{\sigma_Y}\right)^2} = \frac{\sigma_Y^2}{\varepsilon^2} = \frac{1}{n} \cdot \frac{\sigma_X^2}{\varepsilon^2} \tag{15}$$

The last step is because the sample mean itself has mean μ and variance σ^2/n, i.e.,

$$E\left(\overline{X}_n\right) = \frac{1}{n}E\left(X_1 + X_2 + \cdots\right) = \frac{1}{n}\left(E\left(X_1\right) + E\left(X_2\right) + \cdots\right) = \mu \tag{16}$$

and

$$Var\left(\overline{X}_n\right) = \frac{1}{n^2}E\left(X_1 + X_2 + \cdots\right) = \frac{1}{n^2}\left(E\left(X_1\right) + E\left(X_2\right) + \cdots\right) = \frac{\sigma^2}{n} \tag{17}$$

From the scaling property of variances $Var(nX) = \sqrt{n} \cdot Var(X)$.