The Law of Large Numbers and the Central Limit Theorem

Guy Lebanon

October 23, 2006

The Weak Law of Large Numbers (WLLN)

We start with two inequalities that will be used in the proof of the WLLN.

Theorem 1 (Markov Inequality). *Let X be a RV taking only non-negative values with finite mean and variance and $a > 0$. Then*

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

Proof. We prove for the continuous case. A similar proof holds for discrete X.

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)dx = \int_0^{\infty} x f_X(x)dx = \int_0^a x f_X(x)dx + \int_a^{\infty} x f_X(x)dx \geq \int_a^{\infty} x f_X(x)dx$$
$$\geq \int_a^{\infty} a f_X(x)dx = aP(X \geq a)$$

□

Theorem 2 (Chebyshev Inequality). *Let X be a RV with finite mean and variance and $a > 0$. Then*

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

Proof. Apply Markov inequality $P(Z^2 \geq a^2) \leq \frac{\mathsf{E}(Z^2)}{a^2}$ for the RV Z^2 where $Z = |X - E(X)|$:

$$P(|X - E(X)| \geq a) = P(Z \geq a) = P(Z^2 \geq a^2) \leq \frac{\mathsf{Var}(X)}{a^2}$$

□

We use below the abbreviation iid for a sequence of RVs that are interdependent and identically distributed (i.e. all X_i have the same distribution, and in particular the same mean and variance).

Definition 1. *Let $Z_1, Z_2, \ldots$ be iid RVs and Z be a RVs. We say that Z_n converge in probability to Z, denoted $Z_n \to^p Z$ if*

$$\lim_{n\to\infty} P(|Z_n - Z| \leq \epsilon) = 1 \qquad \forall \epsilon > 0.$$

Theorem 3 (The Weak Law of Large Numbers). *Let $X_1, X_2, \ldots$ be a sequence of iid RV with finite expectation μ and variance σ^2 and $\epsilon > 0$. Then empirical means converge in probability to their expectations:*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{p} \mu \qquad or \qquad \lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \geq \epsilon\right) = 0$$

Notice that $\frac{1}{n}\sum_{i=1}^n X_i$ is a RV and the probability above is a number for every n and thus is a sequence of numbers.

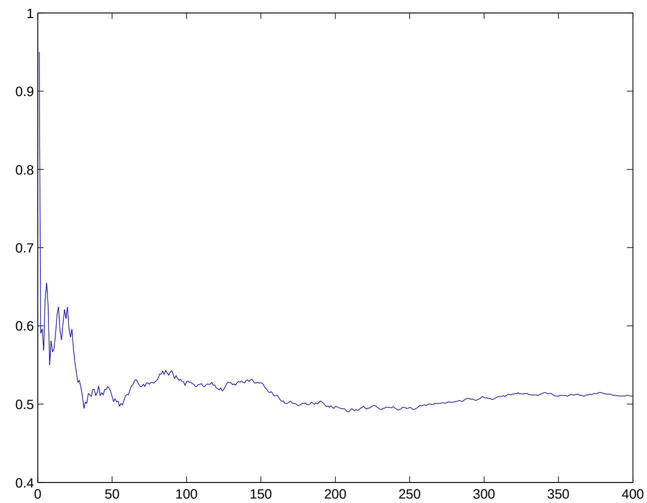

Figure 1: 400 samples $x_1, \ldots, x_{400}$ were randomly generated from a uniform distribution on $[0, 1]$. The curve shows the sample average $\frac{1}{n}\sum_{i=1}^{n} x_i$ as a function of n (varying from 1 to 400). As indicated by the WLLN, the sample average approach the true expectation $\mu = 0.5$ for large n.

Proof. We note that $\mathsf{E}\left(\sum_{i=1}^{n} X_i/n\right) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$ (by X_i identically distributed) and $\mathsf{Var}\left(\sum_{i=1}^{n} X_i/n\right) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \sigma^2/n$ (by X_i independent and identically distributed) and proceed by applying Chebyshev inequality to $\frac{1}{n}\sum_{i=1}^{n} X_i$

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\right| \geq \epsilon\right) \leq \frac{\mathsf{Var}\left(\sum_{i=1}^{n} X_i/n\right)}{\epsilon^2}$$

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) \leq \frac{\frac{1}{n^2}\sum_{i=1}^{n} \sigma^2}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \longrightarrow_{n\to\infty} 0$$

Since the sequence of probabilities are non-negative and its limit is no greater than 0 the theorem is proven. ☐

The WLLN shows that the sequence of RVs $\frac{1}{n}\sum_{i=1}^{n} X_i$ are increasingly centered around the mean μ. In other words, the average over n samples is increasingly likely to be close to μ as $n \to \infty$ (Figure 1). As n increases, the sample average $\frac{1}{n}\sum_{i=1}^{n} x_i$ becomes a good approximation to $\mathsf{E}(X_i) = \mu$.

There is more powerful version of the law of large numbers called the strong law of large numbers. It states that empirical means converge almost surely to their means, denoted $\bar{X}_n \to^{a.s.} \mu$ which formally amounts to

$$P\left(\lim_n \frac{1}{n}\sum_{i=1}^{n} X_n = \mu\right) = 1 \quad \text{or} \quad P\left(\lim_n \left|\frac{1}{n}\sum_{i=1}^{n} X_n - \mu\right| = 0\right) = 1$$

and it implies the weak law of large numbers. An even stronger version is called the uniform strong law of large numbers. It states that if $U(x, \theta)$ is a continuous function in $\theta \in \Theta$ and Θ is compact (closed and bounded subset of $\mathbb{R}^n$) then

$$P\left(\lim_n \sup_{\theta\in\Theta} \left|\frac{1}{n}\sum_{i=1}^{n} U(X_i, \theta) - \mathsf{E}\, U(X_1, \theta)\right| = 0\right) = 1.$$

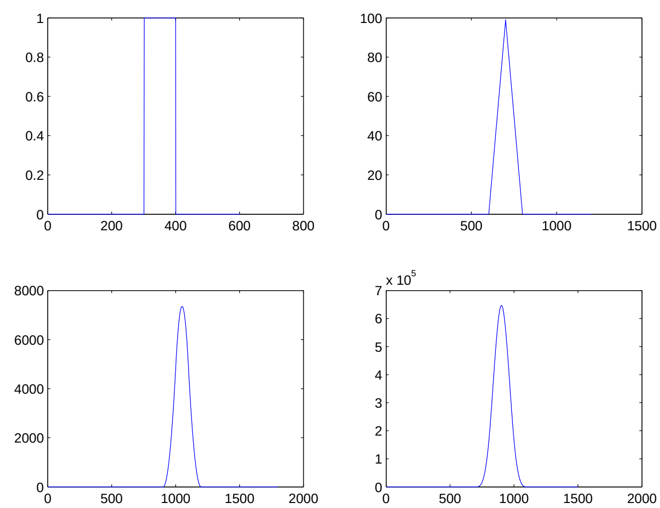

Figure 2: The convolution of the pdf of a uniform distribution (upper left) with itself twice (upper right), three times (bottom left) and four times (bottom right) become increasingly similar to a normal pdf.

The Central Limit Theorem (CLT)

The central limit theorem (CLT) states informally that a sum of many iid RVs will be distributed normally, no matter what is the original distribution of the RVs.

Theorem 4 (Central Limit Theorem). *Let $X_1, X_2, \ldots$ be a sequence of iid RVs with finite mean μ and variance σ^2.*

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^{n}(X_i-\mu)}{\sigma\sqrt{n}} \in A\right) = P(Z \in A) = \int_A \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

where Z is a normal RV with $\mu = 0$ and $\sigma^2 = 1$.

Note that $\frac{\sum_{i=1}^{n}(X_i-\mu)}{\sigma\sqrt{n}}$ is a RV (for every n) with mean 0 and variance 1:

$$\mathsf{E}\left(\frac{\sum_{i=1}^{n}(X_i-\mu)}{\sigma\sqrt{n}}\right) = \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(\mathsf{E}(X_i)-\mu) = 0$$

$$\mathsf{Var}\left(\frac{\sum_{i=1}^{n}(X_i-\mu)}{\sigma\sqrt{n}}\right) = \frac{1}{\sigma^2 n}\sum_{i=1}^{n}\mathsf{Var}(X_i-\mu) = \frac{1}{\sigma^2 n}\sum_{i=1}^{n}\mathsf{Var}(X_i) = \frac{n\sigma^2}{n\sigma^2} = 1.$$

Recall that the pdf of a sum of RVs is given by the convolution of the pdfs. From the CLT we obtain that the convolution of an *arbitrary* function representing the pdf of X with itself many times results in the normal pdf (Figure 2).

The CLT is very useful and surprising. It states that if we are looking for a distribution of a RV Y that is a sum of many other iid RVs $Y = X_1 + \cdots + X_n$ we are guaranteed that its distribution will be close to normal (for large n). This is a standard argument used in many areas for choosing the normal distribution as a model for the data.

The CLT can be used to compute probabilities involving sums of RVs with unknown distribution.

Example (textbook example 5.11): In a restaurant orders $X_1, X_2, \ldots$ are received in an iid fashion with expectation \$8 and variance \$4 (note that the distribution of the orders is unknown). The owner is interested in computing the probability that first 100 orders are greater or equal to 840 $P(\sum_{i=1}^{100} X_i \geq 840)$.

To approximate that quantity using the CLT we need to transform it to a form that enable us to invoke the CLT:

$$P\left(\sum_{i=1}^{100} X_i \geq 840\right) = P\left(\frac{\sum_{i=1}^{100} X_i - 8\cdot 100}{2\cdot 10} \geq \frac{840 - 8\cdot 100}{2\cdot 10}\right) \approx P(Z \geq 2) = \int_2^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

where the last expression is commonly found in tables or by numerical integration.

An equivalent formulation of the CLT is that

$$\lim_{n\to\infty} P\left(\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \in A\right) = P(Z \in A) = \int_A \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

where $\bar{X} = \frac{1}{n}\sum X_i$. Since a $N(0,1)$ RV is almost always within a constant times its standard deviation, we can informally say that for large n, $\bar{X} - \mu \sim c\frac{1}{\sqrt{n}}$ (where $\sim$ here implies on the order of). In other words, for large n, the convergence of $\bar{X}$ to μ is of the order $n^{-1/2}$. This is sometimes written as $\bar{X} = \mu + O_p(n^{-1/2})$ (the O_p notation may be given a precise formulation which generalizes the O notation from calculus).

A multivariate form of the CLT may be expressed as follows. Let $X_1, \ldots, X_n$ be iid random vectors with mean μ and covariance matrix Σ. Then

$$\lim_{n\to\infty} P\left(\sqrt{n}(\bar{X}_n - \mu) \in A\right) = P(Z_\Sigma \in A)$$

where Z_Σ is a multivariate normal RV $Z_\Sigma \sim N(0, \Sigma)$.