ECE 361 Probability for Engineers (Fall, 2016)
Lecture 10b

§5.4 The central limit theorem

Given iid $\{X_1, X_2, \ldots\}$ with $\mathbb{E}[X] = \mu < \infty$ and $\mathrm{var}(X) = \sigma^2 < \infty$, we define $\{M_1, M_2, \ldots\}$ and $\{S_1, S_2, \ldots, \}$ as

$$S_n = \sum_{i=1}^n X_i, \quad M_n = \frac{S_n}{n}. \tag{1}$$

Now

$$\mathbb{E}[M_n] = \mu, \ \mathrm{var}(M_n) = \frac{\sigma^2}{n}, \ \mathbb{E}[S_n] = n\mu, \ \mathrm{var}(S_n) = n\sigma^2. \tag{2}$$

By WLLN, M_n converges in probability to a constant μ, but $S_n \to \infty$. These are two extremes in a sense: M_n converges to a degenerate RV (since the variance converges to 0), while S_n diverges (to ∞) and its variance is unbounded. We would like to have an intermediate regime — it turns out the right scaling is the standardization of S_n. Recall a standardization of a RV, say Y, is the RV $Z = (Y - \mathbb{E}[Y])/\mathrm{std}(Y)$ which has the property $\mathbb{E}[Z] = 0$ and $\mathrm{var}(Z) = 1$. We standardize S_n as

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}, \ n \in \mathbb{N}. \tag{3}$$

It turns out that Z_n converges to the normal distribution, regardless[1] of the the distribution of the underlying $\{X_i\}$. This is the central limit theorem (CLT).

Theorem 1. *Central limit theorem. The CDF of Z_n converges to the standard normal CDF as $n \to \infty$:*

$$\lim_{n\to\infty} \mathbb{P}(Z_n \leq z) = \Phi(z), \ \forall z \in \mathbb{R}, \tag{4}$$

where recall $\Phi(z) = F_Z(z) = \mathbb{P}(Z \leq z)$ for $Z \sim N(0,1)$:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \mathrm{e}^{-\frac{x^2}{2}} \mathrm{d}x. \tag{5}$$

The proof is beyond the scope of this course.

Approximations based on the central limit theorem

The CLT is useful for approximating the distribution of S_n using only knowledge of μ, σ:

$$\mathbb{P}(S_n \leq c) = \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq \frac{c - n\mu}{\sigma\sqrt{n}}\right) = \mathbb{P}\left(Z_n \leq \frac{c - n\mu}{\sigma\sqrt{n}}\right) \approx \mathbb{P}\left(Z \leq \frac{c - n\mu}{\sigma\sqrt{n}}\right) = \Phi\left(\frac{c - n\mu}{\sigma\sqrt{n}}\right). \tag{6}$$

Note the approximation improves as n increases.

Example. Consider 100 packages with weights that are independent RVs uniformly distributed between 5 and 50 pounds. What is the probability the total weight exceeds 3000 pounds? Let $X_1, \ldots, X_{100}$ be the package weights, we want $\mathbb{P}(S_{100} \geq 3000)$. Note $\mu = 27.5$ and $\sigma^2 = 168.75$. Then:

$$\mathbb{P}(S_{100} \leq 3000) \approx \Phi\left(\frac{3000 - 100 \cdot 27.5}{\sqrt{168.75 \cdot 100}}\right) \approx \Phi(1.92) \approx 0.9726. \tag{7}$$

Thus:

$$\mathbb{P}(S_{100} \geq 3000) = 1 - \mathbb{P}(S_{100} \leq 3000) \approx 0.0274. \tag{8}$$

[1]The CLT does require the mean and variance to be finite: $\mu < \infty, \sigma < \infty$.

Example. Processing times for a part are independent RVs uniform on $[1,5]$. We want to find the probability that the number of parts processed within 320 time units, N_{320} is at least 100. Let $X_1,\ldots,X_{100}$ be the processing times of the first 100 parts, and $S_{100} = X_1 + \cdots + X_{100}$ the total processing time of those parts. Then the key observation is:

$$\{N_{320} \geq 100\} = \{S_{100} \leq 320\}. \tag{9}$$

In words, the event that the number of parts processed in the first 320 seconds is at least 100 equals the event that the processing time of the first 100 parts is at most 320 seconds. Now use the CLT approximation:

$$\mathbb{P}(N_{320} \geq 100) = \mathbb{P}(S_{100} \leq 320) = \mathbb{P}\left(\frac{S_{100} - 3\cdot 100}{\sqrt{100\cdot\frac{4}{3}}} \leq \frac{320 - 3\cdot 100}{\sqrt{100\cdot\frac{4}{3}}}\right) \approx \Phi\left(\frac{320 - 3\cdot 100}{\sqrt{100\cdot\frac{4}{3}}}\right) = \Phi(1.73) = 0.9582. \tag{10}$$

Example. Recall the polling problem. We wish to ensure we poll enough people so that the probability of substantial error is sufficiently small, i.e., choose $n(p,\epsilon)$ such that $\mathbb{P}(|M_n - p| \geq \epsilon) \leq \delta$. To that end, observe by symmetry

$$\mathbb{P}(|M_n - p| \geq \epsilon) \approx 2\mathbb{P}(M_n - p \geq \epsilon) = 2\mathbb{P}(S_n \geq n(p+\epsilon)) = 2\mathbb{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \geq \frac{n(p+\epsilon) - np}{\sqrt{np(1-p)}}\right). \tag{11}$$

Now use the CLT:

$$\mathbb{P}(|M_n - p| \geq \epsilon) \approx 2\left(1 - \Phi\left(\sqrt{\frac{n}{p(1-p)}}\epsilon\right)\right) \tag{12}$$

Again, the worst case error is for p is $p = 1/2$ which gives the approximation:

$$\mathbb{P}(|M_n - p| \geq \epsilon) \leq 2\left(1 - \Phi(2\sqrt{n}\epsilon)\right) \tag{13}$$

Fig. 1 compares this approximation with the Chebychev bound. Note the dramatic improvement of the CLT approximation relative to the Chebychev bound – the Chebychev bound is very conservative. If we now solve for n required to ensure the probability of $\epsilon = 1\%$ error is at most $\delta = 5\%$ we find:

$$2\left(1 - \Phi(2\sqrt{n}\epsilon)\right) = \delta \Rightarrow n \geq 9604. \tag{14}$$

Compare with $n \geq 50,000$ required by using the more conservative Chebychev bound. This is a five-fold reduction in the required sample size.

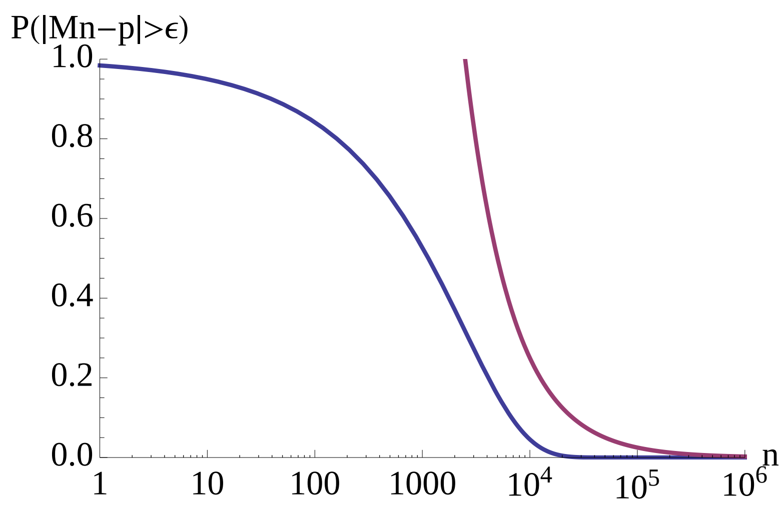

Figure 1: The Chebychev bound $1/(4n\epsilon^2)$ (red) and the CLT bound $2\left(1 - \Phi(2\sqrt{n}\epsilon)\right)$ (blue) on $\mathbb{P}(|M_n - p| \geq \epsilon)$ for $\epsilon = 0.01$. These curves intersect the line $\delta = 0.05$ at $n = 9604$ (CLT) and $n = 50,000$ (Chebychev).

Chernoff's and Hoeffding's inequality

Recall the Markov inequality says that any nonnegative RV X has a tail probability upper bound

$$\mathbb{P}(X > a) \leq \frac{\mathbb{E}[X]}{a}, \ a > 0. \tag{15}$$

Define the *moment generating function* (MGF) with parameter s for a RV X

$$\phi_X(s) \equiv \mathbb{E}[\mathrm{e}^{sX}] = \int_{-\infty}^{\infty} \mathrm{e}^{sx} f_X(x) \mathrm{d}x. \tag{16}$$

Consider the first derivative of the MGF:

$$\phi'_X(s) = \frac{\mathrm{d}}{\mathrm{d}s}\phi_X(s) = \frac{\mathrm{d}}{\mathrm{d}s}\int_{-\infty}^{\infty} \mathrm{e}^{sx} f_X(x)\mathrm{d}x = \int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}s}\mathrm{e}^{sx} f_X(x)\mathrm{d}x = \int_{-\infty}^{\infty} x\mathrm{e}^{sx} f_X(x)\mathrm{d}x = \mathbb{E}[X\mathrm{e}^{sX}]. \tag{17}$$

Observe $\phi'_X(0) = \mathbb{E}[X]$. More generally one can show that $\phi_X^{(k)}(0) = \mathbb{E}[X^k]$, i.e., the kth derivative of the MGF evaluated at zero is the kth moment of X; this explains the name MGF.

The MGF is used for many things in probability; here we simply discuss its use in Chernoff's inequality.

Theorem 2. *(Chernoff bound) For any RV X:*

$$\mathbb{P}(X > a) \leq \inf_{s>0} \mathrm{e}^{-sa}\phi_X(s). \tag{18}$$

The inf stands for "infimum" and may be thought of as a "minimum". The proof of Chernoff's inequality is immediate from Markov's inequality: for $s > 0$

$$\mathbb{P}(X > a) = \mathbb{P}(\mathrm{e}^{sX} > \mathrm{e}^{sa}) \leq \frac{\mathbb{E}[\mathrm{e}^{sX}]}{\mathrm{e}^{sa}} = \mathrm{e}^{-sa}\phi_X(s). \tag{19}$$

As the above argument holds for *any* $s > 0$ it follows that the tightest bound is obtained my minimizing the RHS over all s. Observe that the Markov, Chebychev, and Chernoff bounds require increasingly sophisticated information about the RV X: Markov requires the number $\mathbb{E}[X]$, Chebychev requires the pair of numbers $(\mathbb{E}[X], \mathrm{Var}(X))$, and Chernoff requires the function $\phi_X(s)$.

Our final bound, Hoeffding's inequality, applies to RVs that have a bounded support. We state it without proof.

Theorem 3. *(Hoeffding's inequality) Let $(X_1, \ldots, X_N)$ be independent and identically distributed RVs with $X_i \in [a, b]$ for $i \in [N]$. Let $S_N = X_1 + \cdots + X_N$ be their sum and let $M_N = S_N/N$ be the empirical average. Then*

$$\mathbb{P}(|M_N - \mathbb{E}[M_N]| \geq \epsilon) \leq 2\mathrm{e}^{-\frac{2N\epsilon^2}{(b-a)^2}}, \ \epsilon > 0. \tag{20}$$

Note that Hoeffding's inequality requires only knowledge of the bounded support of each X_i, i.e., $[a, b]$.

As an application, consider the case of estimating the true fraction of voters $p \in (0, 1)$ in favor of a candidate through a random IID sample $(X_1, \ldots, X_N)$, where $X_i \sim \mathrm{Ber}(p)$, and thus $\mathbb{E}[M_N] = \mathbb{E}[X_i] = p$. Here $X_i \in \{0, 1\}$ and thus $a = 0$ and $b = 1$:

$$\mathbb{P}(|M_N - p| \geq \epsilon) \leq 2\mathrm{e}^{-2N\epsilon^2}, \ \epsilon > 0. \tag{21}$$

Using our prior values of $\delta = 0.05$ and $\epsilon = 0.01$ we see that the Chernoff bound assures us that a sufficient number of samples is

$$2\mathrm{e}^{-2N\epsilon^2} = \delta \ \Leftrightarrow \ N = -\frac{\log(\delta/2)}{2\epsilon^2} \approx 18,444. \tag{22}$$

References

[1] *Introduction to Probability, 2nd Edition* by Dimitri P. Bertsekas and John N. Tsitsiklis, Athina Scientific Press, 2008.