EE126: Probability and Random Processes

Lecture 16: Central Limit Theorem

Abhay Parekh

UC Berkeley

March 15, 2011

Limiting Behavior of Random Variables

We observe a sequence $X_1, X_2 \ldots$ iid random variables. Let $M_n = \frac{\sum X_i}{n}$ be the Sample Mean.

1. $E[M_n] = \frac{nE[X_i]}{n} = E[X_i]$
2. Assuming $var(X_i)$ exists, $var(M_n) = \frac{n\ var(X_i)}{n^2} = \frac{var(X_i)}{n}$

As $n \to \infty$: $E[M_n] = E[X_i]$ and $var(M_n) = 0$.
What happens $|X - E[X_i]|$?

Markov Inequality

If X can only take non-negative values then

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

for all $a > 0$.

This inequality makes no assumptions on the existence of variance and so it can't be very strong for typical distributions. In fact, it is quite weak.

Chebyshev Inequality

If X is a random variable with finite mean and variance σ^2, then

$$P(|X - E[X]| \geq c) \leq \frac{\sigma^2}{c^2}$$

for all $c > 0$.
Also, letting $c = k\sigma$:

$$P(|X - E[X]| \geq k\sigma) \leq \frac{1}{k^2}$$

Now we connect variance to probability...If $var(X) = 0$ then

$$P(|X - E[X]| \geq c) = 0$$

for all c.

Chernoff Bound

Also builds on Markov: $P(X \geq a) \leq \frac{E[X]}{a}$ for $a > 0$.
Pick $X = e^{Ys}$. Then

$$P(e^{Ys} \geq a) \leq \frac{M_Y(s)}{a}$$

Now let $a = e^{sb}$. Then for $s > 0$:

$$P(Y \geq b) \leq e^{-sb} M_Y(s)$$

and for $s < 0$:

$$P(Y \leq b) \leq e^{-sb} M_Y(s)$$

Note that the LHS does not depend on s (other than its sign) so we can optimize the RHS to get the best bound.
Since $M_Y(s)$ has all the information contained in $f_Y(y)$ if the distribution has tails that drop off sharply, that will be reflected in our bound.

Weak Law of Large Numbers

It turns out that the WL also holds when variances are infinite, but we can't prove that here. In the Homework!

Weak Law of Large Numbers

If $X_1, ..., X_n$ are iid random variables with mean μ then for every $\epsilon > 0$:

$$P(|\frac{X_1 + \ldots X_n}{n} - \mu| \geq \epsilon) \to 0, \quad \text{as } n \to \infty.$$

This tells us that empirical frequencies are good estimates of p.

What does the Weak Law Really Mean?

WLLN: $\lim_{n\to\infty} P(|M_n - \mu| \geq \epsilon) = 0$.
Just using the defn of limit: For any $\epsilon, \delta > 0$, there exists a number $n(\epsilon, \delta)$ such that

$$P(|M_n - \mu| \geq \epsilon) \leq \delta \quad \text{for all } n \geq n(\epsilon, \delta)$$

- ϵ: (Lack of)Confidence level
- δ: Accuracy level
- $n(\epsilon, \delta)$: threshold function for a given level of confidence and accuracy

What this is saying is that if we compute M_n for large n then:
Almost Always, $|M_n - \mu| < \epsilon$.
We say that M_n **converges to** μ **in probability**.

Convergence in Probability

Let $Y_1, Y_2, \ldots$ be a sequence of random variables. Then we say that the sequence converges to a number a in probability if for every $\epsilon > 0$:

$$\lim_{n \to \infty} P(|Y_n - a| \geq \epsilon) = 0$$

Example: Uniform Distribution

Suppose $X_1, X_2, ..X_n$ are uniformly distributed on $[-1, 1]$:

1. $Y_n = \frac{X_n}{n}$: $Y_n \leq y \Rightarrow X \leq yn$ so
   $F_Y(y) = F_X(yn) \Rightarrow f_Y(y) = nf_X(yn)$.
   Y is uniform over $[-\frac{1}{n}, \frac{1}{n}]$.

   $$P(|Y_n| \geq \epsilon) = 0 \text{ if } n > \frac{1}{\epsilon}$$

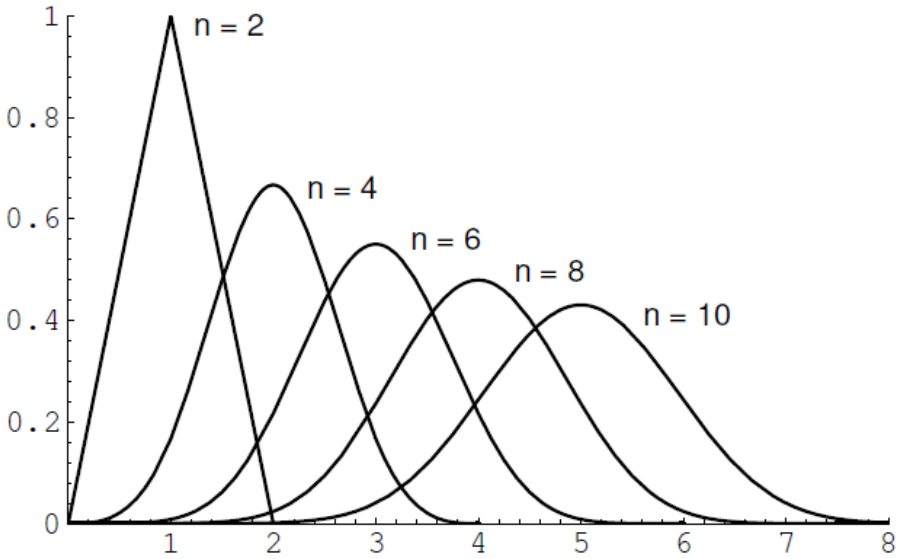
   So Y_n converges to 0 in probability.
2. $Y_n = (X_n)^n$: $P(|Y_n| \geq \epsilon) = P(|X_n^n| \geq \epsilon) = P(|X_n|^n \geq \epsilon) =$
   $P(X_n \geq \epsilon^{\frac{1}{n}}) + P(X_n \leq -\epsilon^{\frac{1}{n}})$. So

   $$P(|Y_n| \geq \epsilon) = 1 - \epsilon^{\frac{1}{n}}$$

   Taking limits on both sides, Y_n converges to 0 in probability.

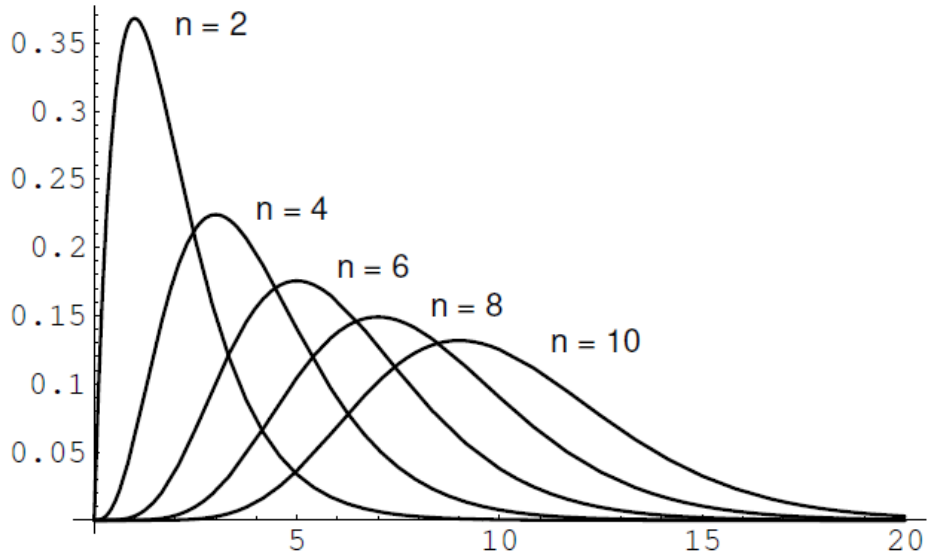
What about the Distribution of S_n?

What happens to $S_n = \sum_i X_i$? Convolving seems to produce symmetry and "bell shapes"...
Example: Uniform on $[0, 1]$:



What about the Distribution of S_n?

What happens to $S_n = \sum_i X_i$? Convolving seems to produce symmetry and "bell shapes"...
Example: Exponential $\lambda = 1$:



Central Limit Theorem

Variance and mean of S_n grow unboundedly with n:
$E[S_n] = n\mu$, $var(S_n) = n\sigma^2 \Rightarrow \sigma_{S_n} = \sqrt{n}\sigma$. So why not look a the normalized version of S_n?
Define

$$\hat{S}_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

($\hat{S}_n$ has mean 0 and variance 1.)
Then, the Central Limit Theorem says that (amazingly):

$$\boxed{\lim_{n\to\infty} P(\hat{S}_n \leq x) = \phi(x), \text{ for every } x}$$

Thus, for large n the random variable $\hat{S}_n$ approximates a standard Normal! I.e.

$$S_n \to N(n\mu, n\sigma^2) \qquad M_n \to N(\mu, \frac{\sigma^2}{n}).$$

Implications of CLT

1. The Distributions of S_n and M_n wipe out all the information in the original information except for μ and σ^2.
2. If there are large number of small and independent factors, the aggregate of these factors will be normally distributed. E.g. Noise.
3. The Gaussian Distribution is very important – many problems involve sums of iid random variables and the only thing one needs to know is the mean and variance.

Simulation

SOCR Sampling Distribution CLT Experiment

Example: Guessing numbers

Alice picks 100 numbers uniformly from $[0, 1]$. Bob must guess the sum within 2. What is the probability he is right if he guesses 55? If X_i is Alice's i^{th} guess, it is difficult to calculate the CDF of $S_{100} = \sum_i X_i$. But, we know that $S_n \approx\sim N(0.5 * 100, 100/12)$. So by the CLT, $\frac{S_{100}-50}{\sqrt{100/12}}$ is approx $N(0, 1)$. Thus

$$P(53 \leq S_{100} \leq 57) \approx \phi(\frac{7}{2.887}) - \phi(\frac{3}{2.887}) = \phi(2.4247) - \phi(1.039)$$

$$= .9925 - .9251 = 0.0649$$

Example: Counting Customers

A store has one cashier who can serve one customer at a time. The time taken to serve a customer $X \sim$ uniform on $[1, 5]$ minutes. What is the probability that at least 200 customers can be served in 8 hours?

We want $P(\sum_{i=1}^{200} X_i \leq 480)$.

$E[S_{200}] = 600$, $var(S_{200}) = 100\ 16/12 = 400/3$.

$$P(\sum_{i=1}^{200} X_i \leq 480) \approx \phi(\frac{480 - 600}{20\sqrt{1/3}}) = \phi(\frac{-6}{\sqrt{3}}) = 1 - \phi(3.464).$$

$$P(\sum_{i=1}^{200} X_i \leq 480) \approx 0.000266$$

Comparing Chebyshev and CLT: Polling

We want to find an estimate of p the prob a randomly chosen voter supports candidate Bob.
We ask 100 randomly sampled voters whether they support him.
$X_i = 1$ if the i^{th} voter says "yes" and $X_i = 0$ otherwise. The X_i are iid. By Chebyshev:

$$P(|M_n - p| \geq 0.1) \leq \frac{p(1-p)}{100\ (0.1)^2} \geq \epsilon) \leq p(1-p)$$

X_i has bounded range $[0, 1]$, and we showed last lecture that $var(X_i) \leq \frac{1}{4}$. So:

$$\boxed{P(|M_n - p| \geq 0.1) \leq \frac{1}{4}}.$$

Now let's apply CLT.

$$P(|M_n - p| \geq 0.1) \approx 2P(M_n - p \geq 0.1)$$

Again, assuming that $var(X_i) = \frac{1}{4}$:

$$\begin{aligned} 2P(M_n - p \geq 0.1) &= 2P(\frac{M_n - p}{\sqrt{(1/4)(1/100)}} \geq (0.1)(20)) \\ &= 2(1 - \phi(2)) = \boxed{0.046} \end{aligned}$$

CLT is much more accurate.

Polling Continued

We ask n randomly sampled voters whether they support Bob. $X_i = 1$ if the i^{th} voter says "yes" and $X_i = 0$ otherwise. The X_i are iid.
We want to be sure with prob ≥ 0.95 that $|M_{100} - p| \leq 0.1$. How many people should we ask?
By Chebyshev:

$$\frac{1}{400n} \leq 0.05 \Rightarrow \boxed{n \geq 50000}$$

By CLT:

$$2(1 - \phi(2 * 0.1 * \sqrt{n})) \leq 0.05$$

$$\phi(2 * 0.1 * \sqrt{n}) \geq 0.975$$

Since $\phi(1.96) = 0.975$:

$$\boxed{n \geq 9604}$$

CLT much better than Chebyshev.

Example: Prediction

Before playing roulette at a casino you watch 100 rounds and count the number of times the result is odd. If the count exceeds k you will decide that the wheel is not fair. Assuming that it is actually fair, what is the probability that you will make a mistake if $k = 55, 60$.

Let X_i be 1 if the round i results in an odd outcome and 0 o.w.

$E[X_i] = 0.5$, $var(X_i) = 0.25$.

$S_{100} = \sum_{i=1}^{100} X_i$ and $S_{100} \approx\sim N(100 * 0.5, 100 * .25)$. I.e. $S_{100} \approx N(50, 25)$. (What is it exactly?)

$$P(S_{100} > k) = P(\frac{S_{100} - 50}{5} > \frac{k - 50}{5}) = 1 - \phi(\frac{k - 50}{5})$$

So

$$P(Mistake) = \begin{cases} 1 - \phi(1) = 0.1587, & \text{if k=55;} \\ 1 - \phi(2) = 0.0228, & \text{if k=60.} \end{cases}$$

De Moivre-Laplace Approximation to Binomial

Suppose S_n is binomial with mean np and variance $np(1-p)$.
We want to find $P(S_n = k)$.
Can't use previous example's approach since there the normal distribution is continuous....Use

$$P(S_n = k) = P(k - \frac{1}{2} \leq S_n \leq k + \frac{1}{2})$$

Now we can approximate using CLT.
Extend this idea to find $P(l \leq S_n \leq k)$ for any non-negative integers l, k.

$$P(k \leq S_n \leq l) \approx \phi(\frac{l + \frac{1}{2} - np}{\sqrt{np(1-p)}}) - \phi(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}})$$

Prediction Re-solved

Before playing roulette at a casino you watch 100 rounds and count the number of times the result is odd. If the count exceeds k you will decide that the wheel is not fair. Assuming that it is actually fair, what is the probability that you will make a mistake if $k = 55, 60$.

$$P(S_{100} > k) = P(S_{100} \geq k+0.5) = P(\frac{S_{100} - 50}{5}) > 1-\phi(\frac{k - 49.5}{5})$$

$$P(Mistake) = \begin{cases} 1 - \phi(1.1) = 0.1357, & \text{if k=55;} \\ 1 - \phi(2.1) = 0.0179, & \text{if k=60.} \end{cases}$$

	Std CLT	De Movire	Exact
k=55	0.1587	0.1357	0.1356
k=60	0.0228	0.0179	0.0108

Proof of the CLT

$X_1, ...X_n$ are iid with mean 0 and variance σ^2. Let $M_X(s)$ the transform of each of the X_i. We know that

$$\hat{S}_n = \frac{\sum_i X_i}{\sqrt{\sigma n}}$$

has mean 0 and variance 1.

$$M_{\hat{S}_n}(s) = E[e^{s\hat{S}_n}] = E[e^{\frac{\sum_i X_i}{\sqrt{\sigma n}}}]$$

Since the X_i are independent:

$$M_{\hat{S}_n}(s) = \Pi_{i=1}^{n} E[e^{\frac{X_i}{\sqrt{\sigma n}}}] = (M_X(\frac{s}{\sigma\sqrt{n}}))^n$$

Proof of the CLT: Part 2

Let's express $M_X(s)$ as the sum of polynomials. To do this we find its Taylor Expansion around $s = 0$. Taylor's Theorem:

$$\boxed{f(x) = \sum_{n=0}^{\infty} \frac{f^n(a)}{n!}(x-a)^n}$$

Expansion of $M_X(s)$ about $s = 0$:

$$M_X(s) = M_X(0) + \frac{M_X'(0)}{1!}(s-0) + \frac{M_X''(0)}{2!}(s-0)^2 + o(s^2)$$

From the properties of transforms: $M_X(0) = 1$,
$M_X'(0) = E[X] = 0$, $M_X''(0) = E[X^2] = \sigma^2$.

$$M_X(s) = 1 + \frac{\sigma^2}{2}s^2 + o(s^2).$$

Thus:

$$M_{\hat{S}_n}(s) = (M_X(\frac{s}{\sigma\sqrt{n}}))^n = (1 + \frac{s^2}{2n} + o(\frac{s^2}{\sigma^2 n}))^n.$$

Proof of the CLT: Part 3

$$M_{\hat{S}_n}(s) = (M_X(\frac{s}{\sigma\sqrt{n}}))^n = (1 + \frac{s^2}{2n} + o(\frac{s^2}{\sigma^2 n}))^n$$

Take limits on both sides as $n \to \infty$:

$$\lim_{n\to\infty} M_{\hat{S}_n}(s) \quad = \quad \lim_{n\to\infty} (1 + \frac{s^2}{2n} + o(\frac{s^2}{\sigma^2 n}))^n$$

Now let's look at $\ln(1 + \frac{s^2}{2n} + o(\frac{s^2}{\sigma^2 n}))^n$ as $n \to \infty$. Since

$$\lim_{x\to 0}(\frac{\ln(1+x) - x}{x}) = 0,$$

it follows that

$$log(1 + x) = x + o(x).$$

$$\begin{aligned} n\ln(1 + \frac{s^2}{2n} + o(\frac{s^2}{\sigma^2 n})) &= n(\frac{s^2}{2n} + o(\frac{s^2}{\sigma^2 n}) + o(\frac{s^2}{2n} + o(\frac{s^2}{\sigma^2 n}))) \\ &= \frac{s^2}{2} + n\ o(\frac{s^2}{\sigma^2 n}) + n\ o(\frac{s^2}{2n}) \end{aligned}$$

Proof of CLT: Part 4

$$\ln(1+\frac{s^2}{2n}+o(\frac{s^2}{\sigma^2 n}))^n = \frac{s^2}{2}+n\ o(\frac{s^2}{\sigma^2 n})+n\ o(\frac{s^2}{2n})$$

But

$$n\ o(\frac{s^2}{\sigma^2 n}) = \frac{s^2}{\sigma^2}\frac{o(s^2/n)}{s^2/n} \to 0 \quad \text{as } n\to\infty$$

and $n\ o(\frac{s^2}{2n}) \to 0$ as well. So

$$\ln(1+\frac{s^2}{2n}+o(\frac{s^2}{\sigma^2 n}))^n \to \frac{s^2}{2}, \quad \text{as } n\to\infty$$

This is the transform for a Standard Normal.
Thus for any x, $F_X(x) \to$CDF of a Standard Normal. I.e., $\hat{S}_n$ **converges in distribution** to $N(0,1)$.

Taylor's Theorem

For any function $f(x)$ which is infinitely differentiable at $x = a$:

$$\int_a^x f'(t)dt = f(x) - f(a) \Rightarrow f(x) = f(a) = \int_a^x f'(t)dt$$

Now integrate by parts to get:

$$\begin{aligned} f(x) &= f(a) + xf'(x) - af'(a) - \int_a^x tf''(t)\,dt \\ &= f(a) + \int_a^x xf''(t)\,dt + xf'(a) - af'(a) - \int_a^x tf''(t)\,dt \\ &= f(a) + (x-a)f'(a) + \int_a^x (x-t)f''(t)\,dt. \end{aligned}$$

The first equation is arrived at by letting $u = f'(t)$ and *dv* = *dt*(in the formula for Integration by parts); the secc

$$\int_a^x xf''(t)\,dt = xf'(x) - xf'(a);$$

the third just factors out some common terms.

Another application yields:

$$f(x) = f(a) + (x-a)f'(a) + \frac{1}{2}(x-a)^2 f''(a) + \frac{1}{2}\int_a^x (x-t)^2 f'''(t)\,dt.$$

By repeating this process, we may derive Taylor's theorem for higher values of *n*.

Strong Law of Large Numbers

Let $X_1, X_2, ...$ be a sequence of iid random variables with mean μ. Then

$$P(\lim_{n\to\infty} \frac{X_1 + ... + X_n}{n} = \mu) = 1.$$

Think of the experiment being repeated n times. Then a basic outcome consists of $(x_1, x_2, ...x_n)$. The Strong Law says that as $n \to \infty$, all the probability is concentrated on those sequences that converge to μ.

We say that M_n **converges to** μ **almost surely**.

What's the difference between the Weak and Strong Laws?

If $X_1, ..., X_n$ are iid random variables with mean μ and $M_n = \frac{1}{n}\sum_i X_i$:

Weak Law of Large Numbers

For every $\epsilon > 0$:

$$\lim_{n\to\infty} P(|M_n - \mu| \geq \epsilon) = 0$$

Strong Law of Large Numbers

$$P(\lim_{n\to\infty} M_n - \mu = 0) = 1.$$

Point of difference: For some $\epsilon > 0$, how often is $|M_n - \mu| > \epsilon$?

1. Weak Law:

$$\frac{\text{Number of times } |M_n - \mu| > \epsilon}{n} \to 0$$

   But this still allows for the numerator to be unbounded with n. (For example $\ln n$.)
2. Strong Law: Has to be a finite number of times for $\lim_{n\to\infty} M_n - \mu$ to be zero.

Example

Suppose time is in discrete units, i.e., $1, 2, ...,$ and $Y_n = 1$ if there is an arrival at time n and $Y_n = 0$ otherwise.
Define $I_k = \{2^k, 2^k + 1, ..., 2^{k+1} - 1\}$ so that:

$$I_1 = \{1\}, \ I_2 = \{2, 3\}, \ I_3 = \{4, 5, 6, 7\} \quad \text{etc.}$$

Suppose that during each interval I_k we have exactly one arrival, and that arrival is equally likely to be at any time in that interval.

$$P(Y_1 = 1) = 1, \ P(Y_2 = 1) = P(Y_3 = 1) = 0.5, \ P(Y_4 = 1) = ... = P(Y_7 = 1) = 0.2$$

Then $P(Y_n = 1) = \frac{1}{2^k}$ if $n \in I_k$.

$$\lim_{n \to \infty} P(Y_n = 1) = \lim_{k \to \infty} \frac{1}{2^k} = 0.$$

Thus, Y_n converges to 0 in probability.
However, $P(\lim_{n \to \infty} Y_n = 0) \neq 0$, since given any finite n there are certain to be an infinite number of arrivals after n.
So Y_n does not converge almost surely.