Approximating by the Normal Distribution

Elizabeth Meckes
AIM / Cornell

March 26, 2007
The central limit theorem (vanilla version)
Let $X_1, X_2, \ldots$ be independent, identically distributed (i.i.d.) random variables with $E[X_i] = 0$, $E[X_i^2] = \sigma^2$.
Let $\tilde{S}_n = \frac{1}{\sigma \sqrt{n}} \sum_{i=1}^{n} X_i$. Then $\forall t \in \mathbb{R}$,
$$
\lim_{n \to \infty} \mathbb{P}(\tilde{S}_n \leq t) = \Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} \, dx.
$$

Key piece of intuition:
One reason that the normal distribution comes up so much is that many random variables are "like" $\tilde{S}_n$: they are sums of a lot of little things, which don't affect each other much.
**Characteristic functions** (a.k.a. Fourier transforms)

**Definition**: If $\mu$ is a probability measure, the characteristic function $\Phi_{\mu}$ of $\mu$ is defined by $\Phi_{\mu}(t) = \int_{-\infty}^{\infty} e^{itx} \, d\mu(x)$.

In terms of r.v.s: if $X$ is a random variable, its characteristic function $\Phi_X$ is defined by $\Phi_X(t) = E[e^{itX}]$.

**The continuity theorem**:

Let $\mu_k$, $\mu$ be probability measures with characteristic functions $\Phi_k$, $\Phi$. Then $\mu_k \xrightarrow{k\to\infty} \mu$ iff $\Phi_k(t) \to \Phi(t)$ for all $t \in \mathbb{R}$.

That is, $\Phi_k(t) \to \Phi(t)$ for every bounded continuous $f: \mathbb{R} \to \mathbb{R}$, $\int f \, d\mu_k \to \int f \, d\mu$. 
Remarks

1. Characteristic functions are one of the most commonly used tools in probability theory and applications.

2. C.f.'s can be used to substantially generalize the CLT; they are particularly well suited to relaxing the assumption that all $\xi X_i$ have the same distributions.

3. C.f.'s can be used to obtain rates of convergence, since the Fourier transform is an isometry on $L^2$.

4. Useful for other limiting distributions, not just Gaussian.
Some basic facts about c.f.s

- $\mathbb{E} [e^{itx}] = \varphi_x(t) = e^{-\frac{t^2}{2}}$ for $x \sim N(0,1)$

- If $X_1, X_2$ are independent random variables, then $\varphi_{X_1+X_2}(t) = \varphi_{X_1}(t) \varphi_{X_2}(t)$.

- If $X$ has finite moments, then
  \[ \varphi_x(t) = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} \mathbb{E}[X^k] \]

  if $X$ has moments up to order $n$, then
  \[ |\varphi_x(t) - \sum_{k=0}^{n} \frac{(it)^k}{k!} \mathbb{E}[X^k]| \]
  \[ \leq \mathbb{E}[\min \left( \frac{|tX|^n}{(n+1)!}, \frac{2|tX|^n}{n!} \right)] \]
Proof of CLT via c.f.s:

Have: \( \{X_i\} \) i.i.d., \( \mathbb{E}[X_i] = 0, \mathbb{E}[X_i^2] = 1 \),
\( S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \). Want: \( S_n \xrightarrow{\mathcal{D}} \mathcal{N} \).

By the continuity theorem, it suffices to show \( \varphi_{\tilde{S}_n}(t) \to \varphi_{\mathcal{N}}(t) = e^{-t^2/2} \quad \forall t \in \mathbb{R} \).

\[
\varphi_{\tilde{S}_n}(t) = \prod_{i=1}^{n} \varphi_{\frac{X_i}{\sqrt{n}}} (t) = \varphi_{\frac{X_i}{\sqrt{n}}} (t) = \varphi_{\frac{X_i}{\sqrt{n}}} \left( \frac{t}{\sqrt{n}} \right).
\]

Now, \( \varphi_{X_i}(0) = \mathbb{E}[e^{i \cdot 0 \cdot X_i}] = 1 \);
\( \varphi_{X_i}'(0) = \mathbb{E}[i \cdot X] = 0 \) by assumption,
\( \varphi_{X_i}''(0) = -\mathbb{E}[\frac{1}{2} X^2] = -\frac{1}{2} \)

\[
\Rightarrow \varphi_{X_i} \left( \frac{t}{\sqrt{n}} \right) = 1 - \frac{t^2}{2n} + o \left( \frac{1}{n} \right)
\]

\[
\Rightarrow \varphi_{X_i} \left( \frac{t}{\sqrt{n}} \right) = \left( 1 - \frac{t^2}{2n} + o \left( \frac{1}{n} \right) \right)^n \xrightarrow{n \to \infty} e^{-t^2/2}.
\]
The Method of Moments

Theorem: Let $Z \sim N(0,1)$ and $\{ X_i \}$ a sequence of random variables such that for each $i$, $X_i$ has moments of all orders. Suppose that

$$E[X^n] \to E[Z^r] = \begin{cases} (r-1)!! & r \text{ even} \\ 0 & r \text{ odd} \end{cases}$$

for each fixed $r \in \mathbb{Z}^+$. Then $X_n \xrightarrow{w} Z$; i.e., $E f(X_n) \to E f(Z)$ for all bounded, continuous $f : \mathbb{R} \to \mathbb{R}$.

Remarks

1. This is how CLT was first proved rigorously (by Chebychev, 1887)
2. Diaconis: "Many young probabilists now shun the method of moments as restricted and heavy-handed. There are some however who realize that moments generally get the job done without taking five years off to develop special theory."

3. It is nearly impossible to get rates of convergence from the method of moments; you can get something with actual equality of the first n moments to those of Z.

4. The theorem above is true for any limiting distribution which is determined by its moments. There are elegant results about when this happens, what sequences of numbers can occur as moments, etc.

5. You must have convergence of all moments for the method to work.
CLT via the method of moments

\( \{X_i\} \) i.i.d., \( \mathbb{E}[X_i] = 0, \mathbb{E}[X_i^2] = 1 \),

assume that \( |X_i| \leq M \) almost surely.

Let \( S_n = X_1 + \cdots + X_n \).

\[
S_n^r = \sum_{p=1}^{r} \sum_{\text{distinct indices}} \frac{r!}{r_1! \cdots r_p!} \frac{1}{p!} \left( \sum_{i_{r_1}, \ldots, i_{r_p}} X_{i_{r_1}}^{r_1} \cdots X_{i_{r_p}}^{r_p} \right)
\]

Define \( A_n(r_1, \ldots, r_p) = \sum_{i_{r_1}, \ldots, i_{r_p}} \left( \frac{1}{n} \right)^{r_1/2} \mathbb{E}[X_{i_{r_1}}] \cdots \mathbb{E}[X_{i_{r_p}}] \).

Then \( \mathbb{E}\left[\left( \frac{1}{\sqrt{n}} S_n \right)^r \right] = \sum_{p=1}^{r} \sum_{\text{distinct indices}} \frac{r!}{r_1! \cdots r_p!} \frac{1}{p!} A_n(r_1, \ldots, r_p) \).

Claim: \( \lim_{n \to \infty} A_n(r_1, \ldots, r_p) = \begin{cases} 1 & r_1 = \cdots = r_p = 2 \\ 0 & \text{otherwise} \end{cases} \)

With the claim, get \( \mathbb{E}\left[\left( \frac{1}{\sqrt{n}} S_n \right)^r \right] \to \frac{r!}{2^{r/2} (r/2)!} = (r-1)!! \quad \text{(r even)} \)

and \( \mathbb{E}\left[\left( \frac{1}{\sqrt{n}} S_n \right)^r \right] \to 0 \), r odd.
To prove the claim:

If any $r_j = 1$, then $\mathbb{E}[X_{i_j}^2] = \mathbb{E}[X_i] = 0$.

$\Rightarrow$ All $r_j \geq 2$. Suppose one $r_j > 2$.

Then $r > 2p$.

$$
\sum' \left( \frac{1}{n} \right)^{r/2} \mathbb{E}[X_{i_1}^2] \cdots \mathbb{E}[X_{i_p}^2]
\leq \frac{M^{r-2p}}{n^{r/2}} \sum' \mathbb{E}[X_{i_1}^2] \cdots \mathbb{E}[X_{i_p}^2]
\leq M^{r-2p} n^{p-r/2} \rightarrow 0^n
$$

Since $p - \frac{r}{2} < 0$.

Finally,

$$
A_n(2, \ldots, 2) = \left( \frac{1}{n} \right)^{r/2} \sum' \mathbb{E}[X_{i_1}^2] \cdots \mathbb{E}[X_{i_p}^2]
= \frac{1}{n^{r/2}} \cdot \left[ n(n-1) \cdots (n - \frac{r}{2} + 1) \right]
\rightarrow 1
$$
The *Lindeberg Method*

The crucial observation here is that if \( Z_1, \ldots, Z_n \) are i.i.d. \( N(0,1) \), then \( Z = \sqrt{n} \sum_{i=1}^{n} \frac{Z_i}{\sqrt{i}} \) is also distributed as \( N(0,1) \).

Idea: to approximate \( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \) by \( Z \), replace the \( X_i \) one at a time by \( Z_i \) and estimate the error.

**Remarks**

1. The approach is fairly bare-hands, and as such has been useful in very general settings (e.g. CLTs in infinite-dimensional settings).

2. Weakening the independence assumption is fairly straightforward here, as is the assumption that all \( X_i \) have the same dist.
CLT via the Lindeberg method

Let \( X_i \) be a collection as usual of i.i.d. random variables, \( \mathbb{E}[X_i] = 0, \mathbb{E}[X_i^2] = 1 \).

Note that \( X_i \) easily satisfy the "Lindeberg condition":

\[
\frac{1}{n} \sum_{i=1}^{n} \int_{\{1 \leq \frac{|X_i|}{\sqrt{n}} < \epsilon^3 \}} x_i^2 \, d\mu(x_i) \to 0 \quad n \to \infty
\]

(here \( \mu \) is the common distribution of the \( X_i \)'s.)

Fix \( f \in \mathcal{C}_c^\infty \): show that

\[
\mathbb{E} f \left( \frac{1}{n} \sum X_i \right) \to \mathbb{E} f (Z) \quad \text{by replacing the } X_i \text{ one at a time by } Z_i.
\]
By Taylor’s theorem, there is a $K$ and $g(h)$ such that $g(h) \leq K \min \{ h^2, h^3 \}$, and

$$|f(x+h_1) - f(x+h_2) - f'(x)(h_1-h_2) - \frac{1}{2} f''(x)(h_1^2-h_2^2)|$$

$$\leq g(h_1) + g(h_2).$$

Define $T_k = X_1 + \cdots + X_{k-1} + Z_{k+1} + \cdots + Z_n$

$\Rightarrow T_n + X_n = S_n$, $T_1 + Z_1 \sim N(0, n)$.

$$|E\left[f\left(\frac{n}{\sqrt{n}}S_n\right) - f(2)\right]|$$

$$\leq \sum_{k=1}^n |E\left[f\left(\frac{T_k + X_k}{\sqrt{n}}\right) - f\left(\frac{T_k + Z_k}{\sqrt{n}}\right)\right]|$$

$$= \sum_{k=1}^n |E\left[f\left(\frac{T_k + X_k}{\sqrt{n}}\right) - f\left(\frac{T_k + Z_k}{\sqrt{n}}\right) - f'(T_n)\left(\frac{X_k - Z_k}{\sqrt{n}}\right)\right] - \frac{1}{2} f''(T_n)\left(\frac{X_k^2 - Z_k^2}{n}\right)|$$

Since $E[X_k] = E[Z_k]$ and $E[X_k^2] = E[Z_k^2]$.
By Taylor, this is
\[
\leq \sum_{k=1}^{n} \left( E\left( \frac{x}{k}\right) + E\left( \frac{\tilde{x}}{k}\right) \right)
\]
\[
= n E\left( \frac{x}{n}\right) + n E\left( \frac{\tilde{\tilde{x}}}{n}\right)
\]

Now,
\[
n E\left( \frac{x}{n}\right) = n \left[ k \int \frac{x}{n} 1_{\{1 \leq x \leq 3\sqrt{n}\}} \, d\mu(x) \right.
\]
\[
+ k \int 1_{\{1 \geq 3\sqrt{n}\}} \, d\mu(x) \right]
\]

We've already seen that the 2nd term goes to zero for fixed \(\varepsilon\), \(n \to \infty\).

\[
n k \int 1_{\{1 \leq 3\sqrt{n}\}} \, d\mu(x) \leq k \cdot n \cdot \varepsilon \cdot \int 1_{\{1 \leq 3\sqrt{n}\}} \, d\mu(x)
\]
\[
\leq k \cdot \varepsilon
\]
The same argument works for \(n E\left( \frac{\tilde{\tilde{x}}}{n}\right)\).
The martingale CLT \( F_{k-1} \subseteq F_k \)

**Definition:** A sequence \( \{X_n, F_n\} \) of r.v.s and \( \sigma \)-algebras is a martingale difference if \( E[X_n | F_{n-1}] = 0 \) almost surely.

(Think \( E[X_n | x_1, \ldots, x_{n-1}] = 0 \) — the expected change in \( X \) at the \( n \)th step is zero)

If \( S_n = \sum_{i=1}^{n} X_i \), this means \( E[S_n | F_{n-1}] = S_{n-1} \)

Idea: drop the independence assumption from the CLT and replace it with this special kind of dependence.
Theorem (Brown): Let \( \{X_n, F_n\} \) be a martingale difference, \( S_n = \sum_{i=1}^{n} X_i \). Define:

\[
\sigma_n^2 = \mathbb{E} [X_n^2 \mid F_{n-1}]
\]

\[
V_n^2 = \sum_{j=1}^{n} \sigma_j^2
\]

\[
S_n^2 = \mathbb{E} [V_n^2] = \mathbb{E} [S_n^2]
\]

Suppose that

- \( V_n^2 S_n^{-2} \to 1 \) in probability as \( n \to \infty \).

- \( S_n^{-2} \sum_{j=1}^{n} \mathbb{E} [X_j^2 \mathbb{1}(|X_j| \geq \varepsilon S_n)] \to 0 \)

in probability as \( n \to \infty \) \( \forall \varepsilon > 0 \) Lindeberg condition

Then:

\[
\lim_{n \to \infty} \mathbb{P} \left[ \frac{S_n}{S_n} \leq t \right] = \Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} \, dx.
\]
Stein's Method

Key observation: the normal distribution is characterized by the fact that \( Z \sim N(0,1) \) if and only if

\[
E \left[ f'(z) - zf(z) \right] = 0
\]

for all \( f \) such that \( E[f'(z)] \) and \( Ezf(z) \) exist.

In fact:

Theorem (Stein): Let \( h: \mathbb{R} \to \mathbb{R} \) be bounded and piecewise \( C^1 \). Then

\[
f(w) = e^{w^2/2} \int_{-\infty}^{w} [h(x) - Eh(z)] e^{-x^2/2} \, dx
\]

is a solution to

\[
f'(w) - wf(w) = h(w) - Eh(z).
\]

Furthermore,

(i) \( \|f\|_{\infty} \leq \sqrt{2\pi} \|h\|_{\infty} \)

(ii) \( \|f'\|_{\infty} \leq 2\|h\|_{\infty} \)

(iii) \( \|f''\|_{\infty} \leq 2\|h'\|_{\infty} \).
Idea: Given $W$ conjectured to be approximately Gaussian, try to bound $E[f'(W) - Wf(W)]$ for a large class of $f$ and use the theorem to show that this implies $E[h(W)] - E[h(Z)]$ is small for a large class of $h$.

Remarks

• This method is very useful in weakening independence assumptions.

• Convergence rates come for free: many notions of distance between distribution take the form

$$d(x, y) = \sup |E_h(x) - E_h(y)|$$

for some class of $h$.

• Any time a distribution can be characterized by a differential (or difference) operator, such an approach may be tried.
Question: How to bound \( E[f''(W) - f''(W')] \)?

Stein's method of exchangeable pairs: Start with \( W \), and make a small (random) change to get \( W' \) so that \((W, W')\) is exchangeable, i.e., \((W, W') \sim (W', W)\).

Example: \( W = \frac{1}{\sqrt{n}} \sum X_i \). Pick \( \{i_1, \ldots, i_n\} \) at random, independent of everything. Let \( W' = W - \frac{1}{n} X_{i_1} + \frac{1}{n} X_{i_1}' \), where \( X_{i_1}' \) is an independent copy of \( X_{i_1} \).

Once you have \( W, W' \), fix \( f \):

\[
0 = E \left[ (W' - W) \left[ f(W') + f(W) \right] \right] \\
= E \left[ (W' - W) \left[ f(W') - f(W) \right] + 2f(W)(W' - W) \right] \\
\geq E \left[ (W' - W)^2 f'(W) + 2f(W)(W' - W) \right] \\
= E \left[ f'(W) \cdot E \left[ (W' - W)^2 |W \right] + 2f(W) \cdot E \left[ W' - W | W \right] \right]
\]
Suppose you've constructed \( W' \) so that
\[
\mathbb{E}[W' - W | W] = -2\lambda W \\
\mathbb{E}[(W' - W)^2 | W] \leq 2\lambda
\]

Then we have:
\[
0 \leq \mathbb{E}[f'(w) \cdot 2\lambda - 2\lambda W f(w)] \\
= 2\lambda \mathbb{E}[f'(w) - W f(w)]
\]

Stein's abstract normal approximation theorem:
Let \((W, W')\) be an exchangeable pair, \(\mathbb{E}W = 0\), \(\mathbb{E}W^2 = 1\), with \(\lambda > 0\) s.t. \(\mathbb{E}[W' - W | W] = -2\lambda W\).

Then if \( h \in C^1 \), \( \|h\|_\infty < \infty \), \( \|W\|_\infty < \infty \),
\[
\left| \mathbb{E}h(W) - \mathbb{E}h(Z) \right| \leq \frac{1}{\lambda} \|h\|_\infty \sqrt{\text{Var} \left( \mathbb{E}[W' - W | W] \right)} \\
+ \frac{1}{4\lambda} \|h''\|_\infty \mathbb{E}(W' - W)^3.
\]
CLT via the method of exchangeable pairs

\[ W = \frac{1}{n} \sum_{i=1}^{n} X_i , \quad W' = W - \frac{1}{n} X_I + \frac{1}{n} X'_I, \]

where \( I \in \{1, \ldots, n\} \) is random and \( X'_i \) is an independent copy of \( X_i \) for each \( i \).

\[ E[W' - W] = \frac{1}{n} E[(X'_I - X_I) W] \]

\[ = -\frac{1}{n} E\left[ \frac{1}{n} \sum_{i=1}^{n} X_i - X_I \right] \]

\[ = -\frac{1}{n} W \]

\[ E[(W' - W)^2] = \frac{1}{n} E\left[ (X'_I)^2 - 2X'_I X_I + X_I^2 \right] \]

\[ = \frac{1}{n^2} \sum_{i=1}^{n} E\left[ (X_i)^2 - 2X_i X_I + X_I^2 \right] \]

\[ = \frac{1}{n} + \frac{1}{n^2} \sum_{i=1}^{n} E[X_i^2] \]

\[ = 2 \frac{1}{n} + \frac{1}{n^2} \sum_{i=1}^{n} E[X_i^2 - 1] \]
\[ \Rightarrow \mathbb{E} \left[ \mathbb{E} \left[ (w' - w)^2 \mid W \right] \right] = \frac{2}{n} \]

\[ \text{Var} \left( \mathbb{E} \left[ (w' - w)^2 \mid W \right] \right) \]
\[ = \mathbb{E} \left[ \frac{1}{n^2} \mathbb{E} \left[ \sum_{i} (x_i - 1)^2 \right] \right] \]
\[ = \frac{1}{n^2} \mathbb{E} \left[ \sum_{i} \sum_{j \neq i} (x_i - 1)(x_j - 1) \right] \]
\[ = \frac{1}{n^3} \mathbb{E} \left[ (x_i - 1)^2 \right] = \mathbb{E} \left[ \frac{x_i^2 - 1}{n^3} \right] \]

\[ \mathbb{E} \left[ (w' - w)^3 \right] = \mathbb{E} \left[ (x_i' - x_i)^3 \right] \cdot \frac{1}{n^{3/2}} \]
\[ \leq \frac{3\sqrt{2}}{n^{3/2}} \mathbb{E} |x_i'|^3 = \frac{3\sqrt{2}}{n^{3/2}} \mathbb{E} |x_1|^3 \]

\[ \Rightarrow |\mathbb{E} h(w) - \mathbb{E} h(z)| \leq \frac{1}{n^{3/2}} \|h\|_{\infty} \mathbb{E} [x_i^2 - 1] \]
\[ + \frac{1}{n^{3/2}} \|h\|_{\infty} \left( \frac{3\sqrt{2}}{4} \mathbb{E} |x_1|^3 \right) \]
Other approaches to Stein's method

Zero-bias coupling:
Given $W$, make $W^*$ close to $W$ such that $\mathbb{E}[f'(W^*)] = \mathbb{E}[Wf(W)]$

Theorem (Goldstein/Reinert)
Let $(W, W^*)$ be defined on a joint probability space with $\mathbb{E}W = 0$, $\mathbb{E}W^2 = 1$ and $W^*$ defined as above. Then for $h \in C^4$,

$$|\mathbb{E} h(W) - \mathbb{E} h(Z)| \leq \frac{1}{3} |h^{(3)}|_{\infty} \sqrt{\mathbb{E}(\mathbb{E}[W^2 - W]^2)} + \frac{1}{8} |h^{(4)}|_{\infty} \mathbb{E}(W^* - W)^2.$$
Size-bias coupling:

**Theorem (Goldstein/Rinott):**

Let $W \geq 0$ be a random variable with $\mathbb{E} W = \lambda$, $\text{Var} W = \sigma^2$, and define $W^*$ by $\mathbb{E}[Wg(W)] = \lambda \mathbb{E}g(W^*)$ for any $g$ s.t. the LHS exists. Then for $h \in C^1_c$,

$$|\mathbb{E}h\left(\frac{W-\lambda}{\sigma}\right) - \mathbb{E}h(Z)| \leq 2\|h\|_\infty \frac{\lambda}{\sigma^2} \sqrt{\text{Var}(\mathbb{E}[W^*-W])} + \|h\|_\infty \frac{\lambda}{\sigma^3} \mathbb{E}(W^*-W)^2.$$
Dependency graphs:

Theorem (Stein): Let $Y_1, \ldots, Y_n$ be r.v.'s, $M_i \in \mathcal{M}_1, \ldots, \mathcal{M}_n$ s.t.

1. $j \in M_i$ iff $i \in M_j$

2. $(Y_i, Y_j)$ is independent of $\{Y_k \mid k \not\in M_i \cup M_j\}$.

Suppose $\mathbb{E}Y_i = 0 \ \forall i$, $\mathbb{E}Y_i^2 = 1 \ \forall i$. Then for $h \in C'_c$,

$$\left| \mathbb{E} h\left(\frac{1}{n} \sum_{i=1}^{n} Y_i\right) - \mathbb{E} h(2) \right|$$

$$\leq \frac{2}{n} \| h \|_{\infty} \sqrt{\mathbb{E} \left\{ \sum_{i=0}^{n} \left( \sum_{j \in M_i} (Y_i Y_j - \mathbb{E}(Y_i Y_j)) \right)^2 \right\}^2}$$

$$+ \frac{1}{n^{1/2}} \| h \|_{\infty} \mathbb{E} \left[ \sum_{i=0}^{n} Y_i \left( \sum_{j \in M_i} Y_j \right)^2 \right].$$