5.4 The Central Limit Theorem

Let $X_1, X_2, \ldots$ be a sequence of IID random variables with $E(X_i) = \mu$, and $\text{var}X_i = \sigma^2 < \infty$, and define

$$Z_n = \frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$$

Then, the **CDF** of Z_n converges to the standard normal CDF

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx$$

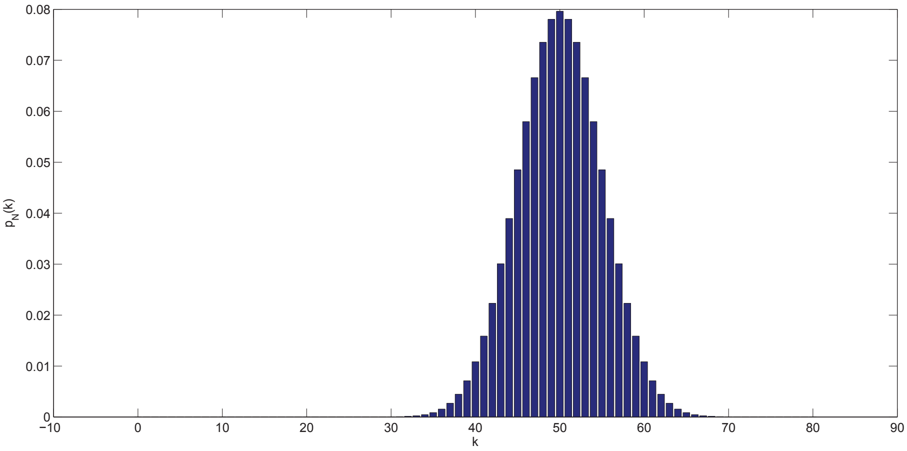
in the sense that

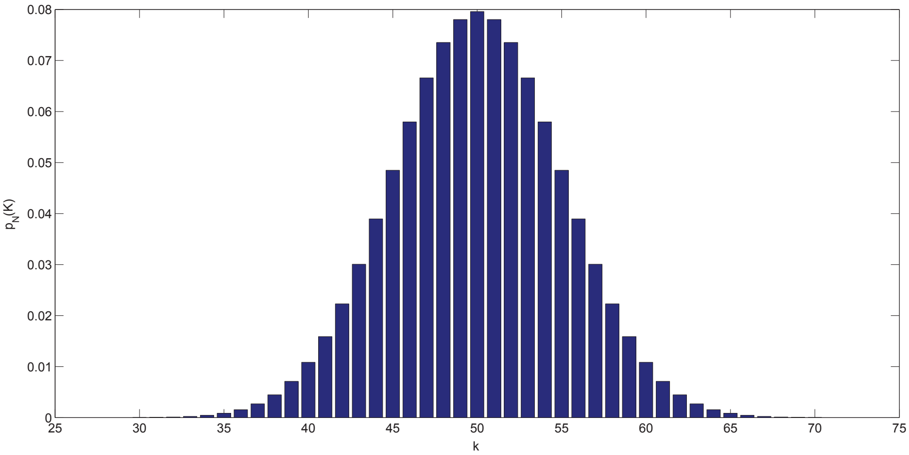
$$\lim_{n\to\infty} \mathrm{P}(Z_n \leq z) = \Phi(z)$$

for every z.

Let's see a very simple application of this that illustrates many of the issues.

Ex: Toss a fair coin 100 times. Find an approximation for the number of H's being between 45 and 55 (inclusive).


$p_N(k)$
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
−10
0
10
20
30
40
50
60
70
80
90
k


$p_N(K)$
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
25
30
35
40
45
50
55
60
65
70
75
k

Simple Proof of the Central Limit Theorem: We will use Transforms. For simplicity, assume $E(X) = 0$, $\sigma = 1$. Consider the transform of Z_n, $\Phi_{Z_n}(s)$.

$$\begin{aligned}
\Phi_{Z_n}(s) &= E[e^{s\frac{X_1+X_2+\ldots+X_n}{\sqrt{n}}}] \\
&= E[e^{s\frac{X_1}{\sqrt{n}}}e^{s\frac{X_2}{\sqrt{n}}}\ldots e^{s\frac{X_n}{\sqrt{n}}}] \\
&= \prod_{i=1}^{n} E[e^{s\frac{X_i}{\sqrt{n}}}] \\
&= (E[e^{s\frac{X_i}{\sqrt{n}}}])^n \\
&= (E[1+\frac{sX_i}{\sqrt{n}}+\frac{(sX_i)^2}{2n}\ldots])^n(\text{ using Taylor Series expansion}) \\
&= (E[1+\frac{(sX_i)^2}{2n}\ldots])^n(\text{ zero mean}) \\
&\to e^{s^2/2}
\end{aligned}$$

Since the transform of Φ_{Z_n} converges to that of a standard Gaussian, and there is a one-to-one mapping between transforms and distributions, we conclude that Φ_{Z_n} converges to a standard Gaussian.

There are other versions of the central limit theorem (CLT), but this basic version assumes that the random variables are independent, identically distributed, with finite mean and variance. Aside from these, there is no requirement on the distribution of X, it can be continuous, discrete, or mixed. One thing to be careful about while using the CLT is that this is a "central" property, i.e. it works well around the mean of the distribution, but not necessarily at the tails (consider, for example, X_i being strictly positive and discrete- then Z_n has absolutely no probability mass below zero, whereas the Gaussian distribution does.)

We shall end this topic with an application of the CLT to the polling problem.

Ex: Revisit the polling problem. Find approximately how large n should be such that the probability that the poll result differs from the true voter fraction in absolute value by more than 0.01 is less than 5 percent. (Answer: $n \geq 10,000$ will suffice.)