Suppose a drunkard is standing at time zero (say 11:00 PM) at some point, and every second he either moves one step to the right, or one step to the left, with equal probability, of where he is at that time. What is the probability that after two minutes, he will be ten or more steps away from where he started? Note that the drunkard will take 120 steps in 2 minutes.

Let the drunkard's movement at the ith step be denoted as X_i; then, $P(X_i = \pm 1) = .5$. So, we can think of X_i as $X_i = 2Y_i - 1$, where $Y_i \sim \text{Ber}(.5), 1 \leq i \leq n = 120$. If we assume that the drunkard's successive movements $X_1, X_2, \ldots$ are independent, then $Y_1, Y_2, \ldots$ are also independent, and so, $S_n = Y_1 + Y_2 + \cdots Y_n \sim \text{Bin}(n, .5)$. Furthermore,

$$|X_1 + X_2 + \cdots + X_n| \geq 10 \Leftrightarrow |2(Y_1 + Y_2 + \cdots + Y_n) - n| \geq 10.$$

So, we want to find

$$\begin{aligned}
&P(|2(Y_1 + Y_2 + \cdots + Y_n) - n| \geq 10) \\
&= P\left(S_n - \frac{n}{2} \geq 5\right) + P\left(S_n - \frac{n}{2} \leq -5\right) \\
&= P\left(\frac{S_n - \frac{n}{2}}{\sqrt{.25n}} \geq \frac{5}{\sqrt{.25n}}\right) + P\left(\frac{S_n - \frac{n}{2}}{\sqrt{.25n}} \leq -\frac{5}{\sqrt{.25n}}\right).
\end{aligned}$$

Using the normal approximation, this is approximately equal to $2[1 - \Phi(\frac{5}{\sqrt{.25n}})] = 2[1 - \Phi(.91)] = 2(1 - .8186) = .3628$.

We present four simulated walks of this drunkard in Fig. 1.6 over a two-minute interval consisting of 120 steps. The different simulations show that the drunkard's random walk could evolve in different ways.

1.17.1 Binomial Confidence Interval

The normal approximation to the binomial distribution forms the basis for most of the confidence intervals for the parameter p in common use. We describe two of these in this section, the *Wald confidence interval* and the *score confidence interval* for p. The Wald interval used to be the textbook interval, but the score interval is gaining in popularity due to recent research establishing unacceptably poor properties of the Wald interval. The derivation of each interval is sketched below.

Let $X \sim \text{Bin}(n, p)$. By the normal approximation to the $\text{Bin}(n, p)$ distribution, for large n, $X \approx N(np, np(1-p))$, and therefore, the standardized binomial variable $\frac{X - np}{\sqrt{np(1-p)}} \approx N(0, 1)$. This implies

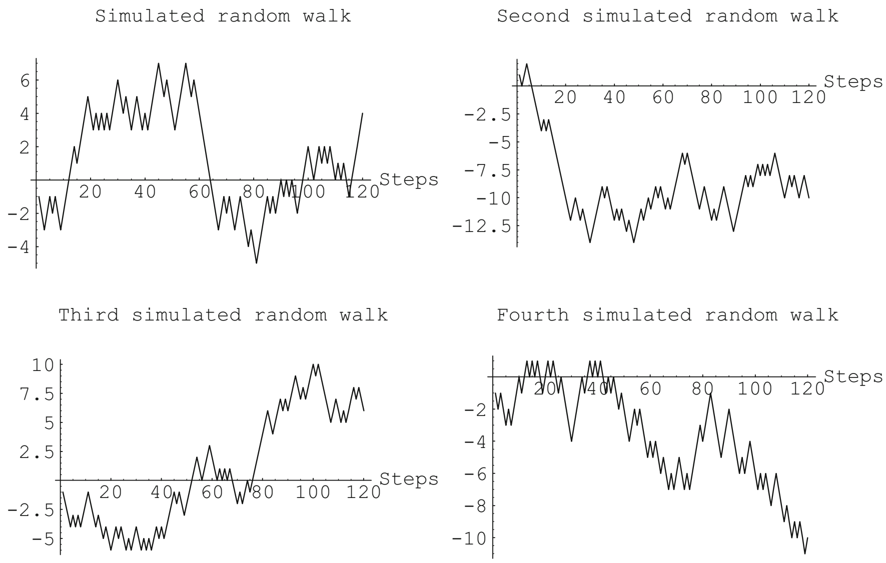

Fig. 1.6 Four simulated random walks

$$
\begin{aligned}
& P\left(-z_{\frac{\alpha}{2}} \leq \frac{X-np}{\sqrt{np(1-p)}} \leq z_{\frac{\alpha}{2}}\right) \approx 1-\alpha \\
\Rightarrow\ & P\left(-z_{\frac{\alpha}{2}}\sqrt{np(1-p)} \leq X-np \leq z_{\frac{\alpha}{2}}\sqrt{np(1-p)}\right) \approx 1-\alpha \\
\Rightarrow\ & P\left(-z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \leq \frac{X}{n}-p \leq z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) \approx 1-\alpha \\
\Rightarrow\ & P\left(\frac{X}{n}-z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \frac{X}{n}+z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) \approx 1-\alpha.
\end{aligned}
$$

This last probability statement almost looks like a confidence statement on the parameter p, but not quite, because $\sqrt{\frac{p(1-p)}{n}}$ is not computable. So, we cannot use $\frac{X}{n} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$ as a confidence interval for p. We remedy this by substituting $\hat{p} = \frac{X}{n}$ in $\sqrt{\frac{p(1-p)}{n}}$, to finally result in the confidence interval

$$\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

This is *the Wald confidence interval* for p.

An alternative and much better confidence interval for p can be constructed by manipulating the normal approximation of the binomial in a different way. The steps proceed as follows. Writing once again $\hat{p}$ for $\frac{X}{n}$,

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\frac{\alpha}{2}}\right) \approx 1-\alpha$$

$$\Rightarrow P\left((\hat{p}-p)^2 \leq z^2_{\frac{\alpha}{2}} \frac{p(1-p)}{n}\right) \approx 1-\alpha$$

$$\Rightarrow P\left(p^2\left(1+\frac{z^2_{\frac{\alpha}{2}}}{n}\right) - p\left(2\hat{p}+\frac{z^2_{\frac{\alpha}{2}}}{n}\right) + \hat{p}^2 \leq 0\right) \approx 1-\alpha.$$

Now the quadratic equation

$$p^2\left(1+\frac{z^2_{\frac{\alpha}{2}}}{n}\right) - p\left(2\hat{p}+\frac{z^2_{\frac{\alpha}{2}}}{n}\right) + \hat{p}^2 = 0$$

has the two real roots

$$p = p_{\pm} = \frac{\hat{p}+\frac{z^2_{\frac{\alpha}{2}}}{2n}}{1+\frac{z^2_{\frac{\alpha}{2}}}{n}} \pm \frac{z_{\frac{\alpha}{2}}\sqrt{n}}{n+z^2_{\frac{\alpha}{2}}}\sqrt{\hat{p}(1-\hat{p})+\frac{z^2_{\frac{\alpha}{2}}}{4n}}.$$

This is the *score confidence interval* for p. It is established theoretically and empirically in Brown, Cai, and DasGupta (2001, 2002) that the score confidence interval performs much better than the Wald interval, even for very large n.

1.17.2 Error of the CLT

A famous theorem in probability places an upper bound on the error of the normal approximation in the central limit theorem. If we make this upper bound itself small, then we can be confident that the normal approximation will be accurate. This upper bound on the error of the normal approximation is known as the *Berry–Esseen bound*. Specialized to the binomial case, it says the following; a proof can be seen in Bhattacharya and Rao (1986) or in Feller (1968). The general Berry–Esseen bound is treated in this text in Chapter 8.

Theorem 1.59 (Berry–Esseen Bound for Normal Approximation). *Let* $X \sim \text{Bin}(n, p)$*, and let* $Y \sim N(np, np(1-p))$*. Then for any real number* x*,*

$$|P(X \leq x) - P(Y \leq x)| \leq \frac{4}{5}\frac{1-2p(1-p)}{\sqrt{np(1-p)}}.$$