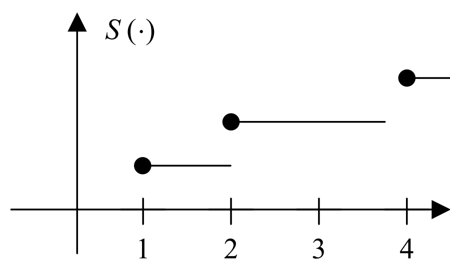
SCALING and CENTERING of DATA
Statistical Regularity

Bernoulli $\Rightarrow$ Brownian Motion

Here: Visual + intuitive convergence.
(Rigorous: Weak convergence, or convergence of stochastic processes *in distribution*).

Bernoulli process: Start with

$$\begin{aligned} \Delta_n &= 2 \quad \text{wp } 1/2 \qquad n = 1, 2, \ldots, \; iid \\ &= 0 \quad \text{wp } 1/2 \; . \\ \text{Define} \qquad S_n &= \Delta_1 + \Delta_2 + \cdots + \Delta_n \,, \quad n \geq 1 \,, \\ S(t) &= \sum_{k=1}^{\lfloor t \rfloor} \Delta_k \quad (= S_{\lfloor t \rfloor}) \;, \quad t \geq 0 : \end{aligned}$$



- Fit $\{S(t), 0 \leq t \leq n\}$, n large, into a 1×1 frame (a computer screen).

What to expect?

SLLN (Strong Law of Large Numbers) $\frac{1}{n} S(n) \xrightarrow{wp1} E\Delta_1 = 1, \; n \uparrow \infty.$

1. Rescaling time: $\{S(nt), \; 0 \leq t \leq 1\}$; at $t = 1 : \; ES(n) \approx n$;

2. Aggregating space: $\{\frac{1}{n} S(nt), \; t \geq 0\} \xrightarrow{wp1} \{t, \; t \geq 0\}, \; n \uparrow \infty.$

CLT (Central Limit Theorem) $\sqrt{n}[\frac{1}{n} S(n) - 1] \stackrel{d}{\Longrightarrow} N(0, 1), \; n \uparrow \infty.$

3. Centering: $\frac{1}{n} S(nt) - t$: deviations from trend.

4. Amplifying: $\sqrt{n}[\frac{1}{n} S(nt) - t] \stackrel{d}{\Longrightarrow} N(0, t), \; n \uparrow \infty.$

Obtained: $S(nt) \approx n[t + \frac{1}{\sqrt{n}} N(0,t)] = nt + \sqrt{n} N(0,t)$
$$\stackrel{d}{=} nt + N(0, nt)$$

Expect $S(t) \stackrel{d}{\approx} t + B(t), \quad t \geq 0,$

where $B = \{B(t),\ t \geq 0\}$ is a Levy process with normal Gaussian marginals, namely a Brownian Motion $BM(0,1) = SBM$ (**S**tandard **B**rownian **M**otion).

Invariance Principle

Let $\Delta_1, \Delta_2, \ldots$ be *iid* with mean $= \mu$ and finite variance $= \sigma^2$ (*general* distribution).

For $S = \{S(t),\ t \geq 0\}$ as before, we have

$$S(t) \stackrel{d}{\approx} \mu \cdot t + \sigma B(t), \quad t \geq 0\,,$$

where $B = \{B(t),\ t \geq 0\}$ is an SBM.

Equivalently, $S \stackrel{d}{\approx} BM(\mu, \sigma^2)$.

Terminology:
$\bar{S}(t) = \mu t$ *fluid approximation*;
$\hat{S}(t) = \sigma B(t)$ *diffusion refinement*.

- **FSLLN** (Functional SLLN)

$$\frac{1}{n} S(nt) \longrightarrow \mu t, \quad t \geq 0, \qquad \text{(convergence wp1, u.o.c.)}$$

(fluid limit)

- **FCLT** (Donsker's Theorem)

$$\sqrt{n}\left[\frac{1}{n} S(nt) - \mu t\right] \Longrightarrow \sigma B(t), \quad t \geq 0, \qquad \text{(weak convergence = convergence in distribution).}$$

(diffusion limit)

Alternatively: Fluid and Diffusion Approximations / Models.

- **Strong Approximations**: imply both FSLLN and FCLT, at the cost of higher moments.

FLUID and DIFFUSION Approximations

Donsker's Theorem in Pictures

$$\Delta_n = \begin{cases} 2 & \text{wp} \quad 1/2 \\ 0 & \phantom{\text{wp}} \quad 1/2 \end{cases} \quad ; \quad S_n = \sum_{k=1}^{n} \Delta_k \quad , \quad S(t) = \sum_{k=1}^{\lfloor t \rfloor} \Delta_k, \quad t \geq 0\,.$$

Fluid approximation: $S(t) \approx t, \ t \geq 0 \quad ; \quad$ supported by

FSLLN $\frac{1}{n} S(nt) \xrightarrow{wp1} t, \quad t \geq 0\,.$

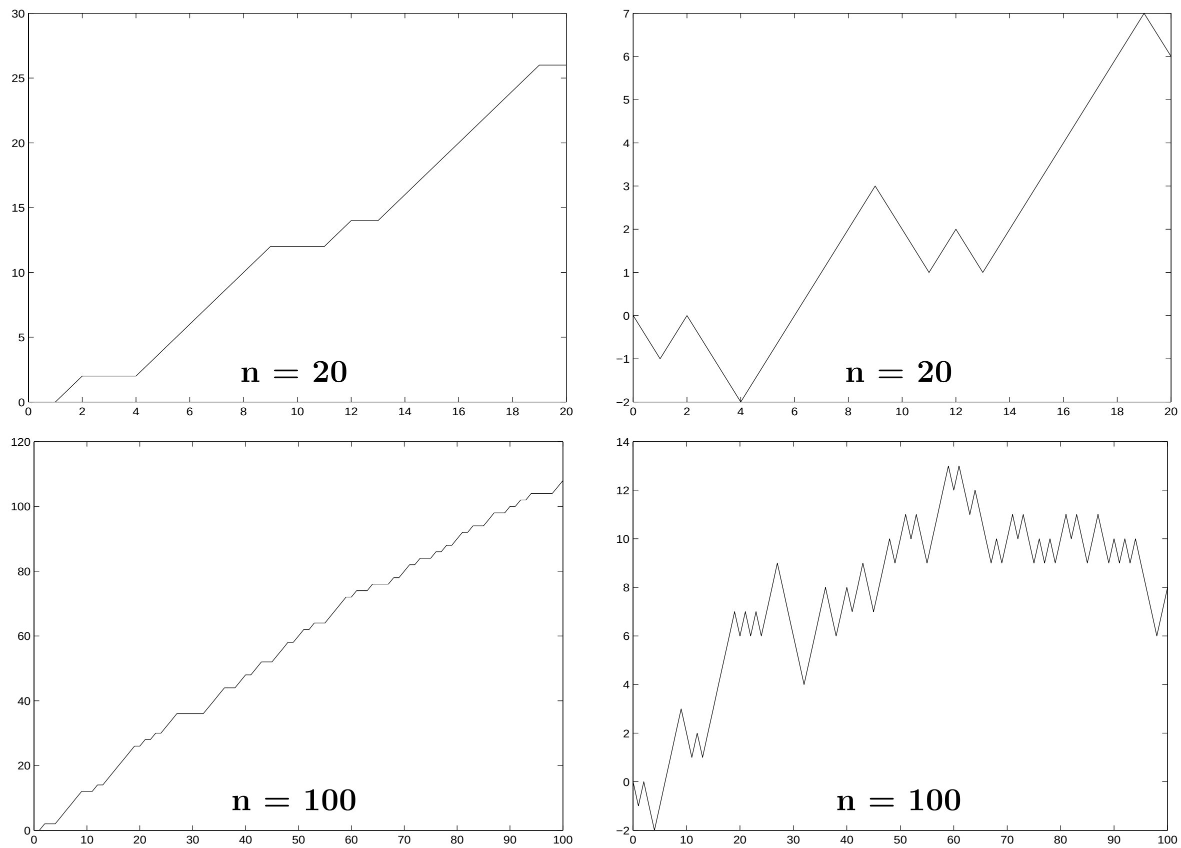
Diffusion refinement: $S(t) \approx t + B(t), \quad t \geq 0,$ supported by

FCLT $\sqrt{n}\left[\frac{1}{n} S(nt) - t\right] \stackrel{d}{\Longrightarrow} B(t), \quad t \geq 0\,,$

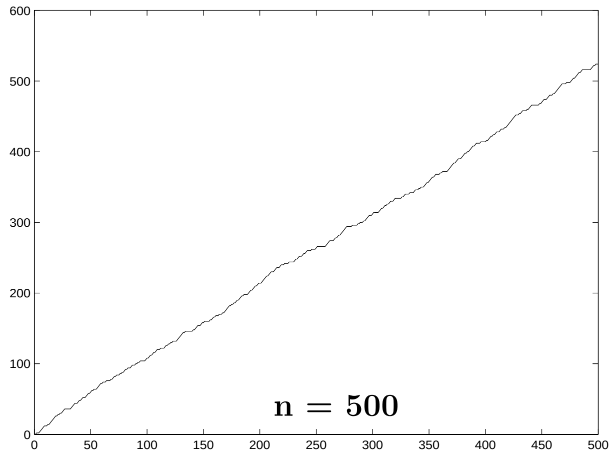
where $B = \{B(t),\ t \geq 0\}$ is Levy with normal (Gaussian) marginals, indeed SBM.

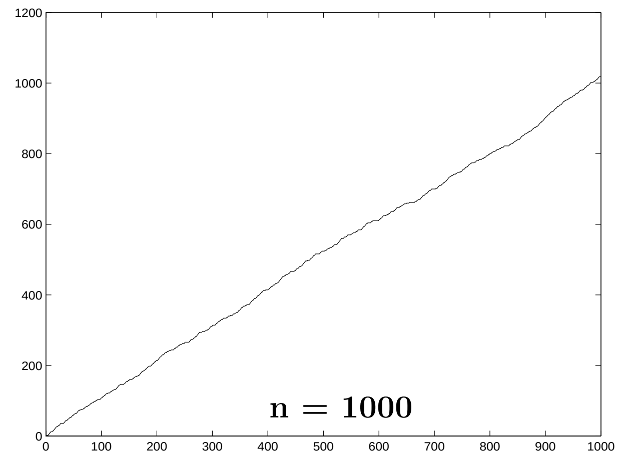
$\square$

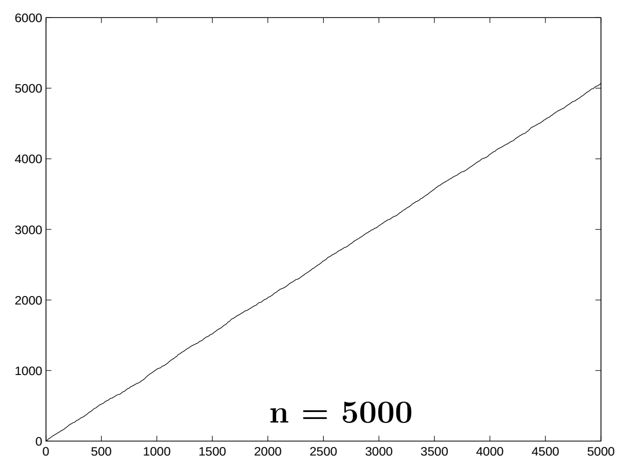
Invariance Principle: $\{\Delta_k\}$ mean μ, std. σ, then $\boxed{S(t) \approx \mu t + \sigma B(t),\ t \geq 0.}$



Fluid approximation: Deterministic Functional
Strong Law of Large Numbers (FSLLN)

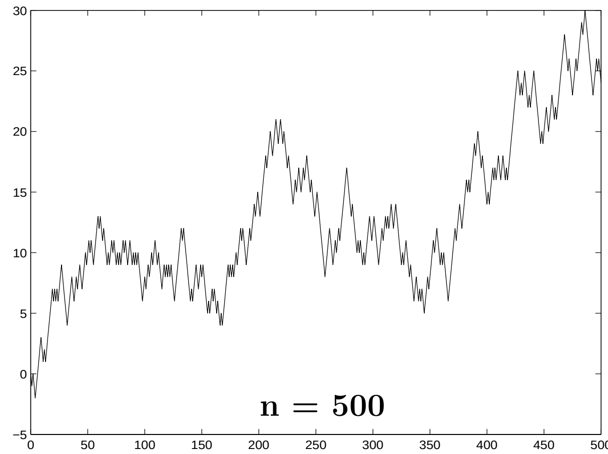


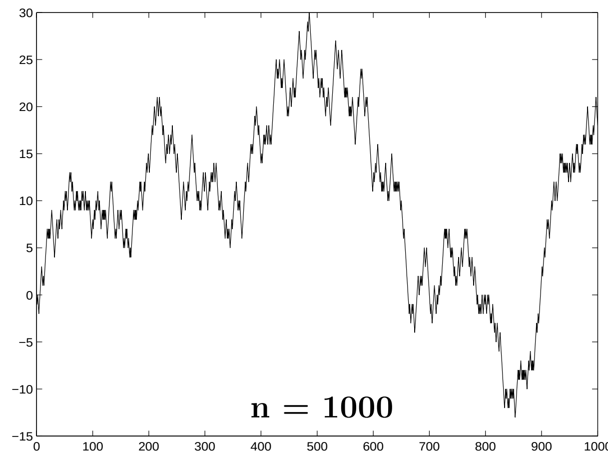


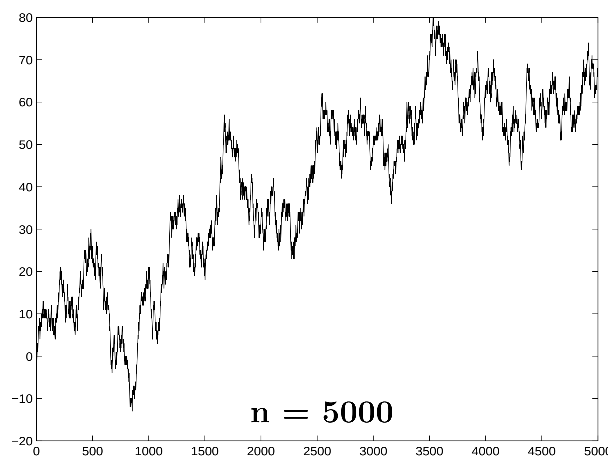


Interpolating $S(0), S(1), \ldots, S(n)$

Diffusion deviation: Stochastic Functional
Central Limit Theorem (FCLT)







Interpolating deviations from trend

Poisson: Via a *sequence* of Bernoulli processes.

Let

$$S^n(t) = \sum_{k=1}^{\lfloor t \rfloor} \Delta_k^n \,, \quad \text{where } \Delta_k^n \begin{array}{ccl} = & 1 & \text{wp } p_n = \dfrac{\lambda}{n} \\ = & 0 & \text{wp } q_n = 1 - p_n \,. \end{array}$$

Plot $\{S^n(t),\ 0 \le t \le n\}$, n large, as before.

What to expect?

Law of **Rare Events** $\quad S^n(n) \stackrel{d}{=} \text{Binomial}(n, \frac{\lambda}{n}) \quad \Rightarrow \quad \text{Poisson } (\lambda).$

Rescaling time: $\{S^n(nt),\ 0 \le t \le 1\}$

$\quad\quad ''$

$\quad\quad \text{Bin } (\lfloor nt \rfloor, \frac{\lambda}{n}) \Rightarrow \text{Poisson } (\lambda t),\ t \ge 0.$

Expect $\quad\quad S^n(nt) \stackrel{d}{\approx} A(t),\ t \ge 0\,,$

where $A = \{A(t),\ t \ge 0\}$ is a Levy process with Poisson marginals, namely a Poisson (λ) process.

Brownian: $S(t) = \sum_{k=1}^{\lfloor t \rfloor} \Delta_k$, $\quad$ where $\Delta_k = 1$ wp p
$\quad\quad 0 \quad q = 1 - p$

FSLLN: $\quad \frac{1}{n} S(nt) \xrightarrow{wp1} p \cdot t$, since $E\Delta_1 = p$.

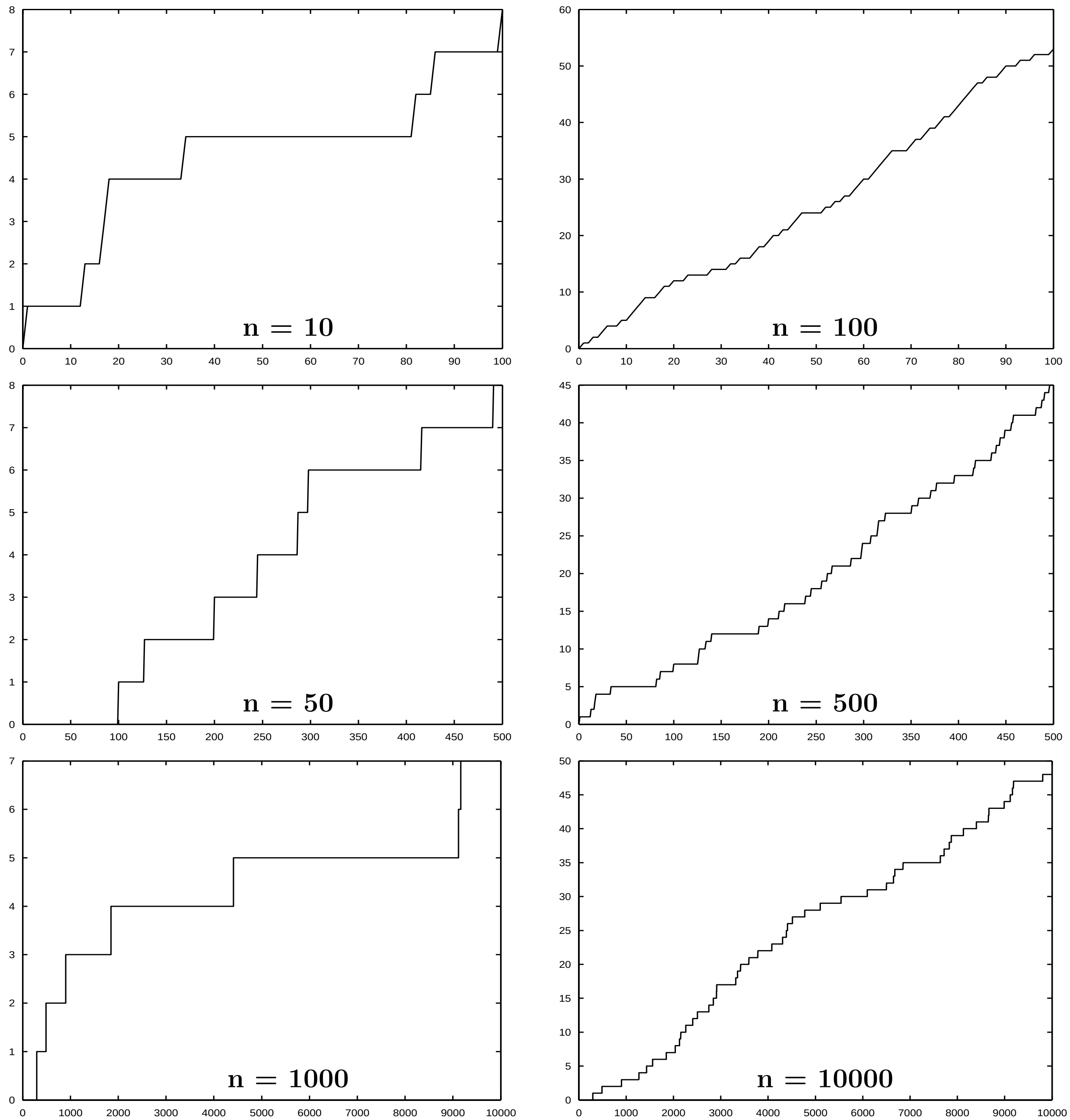
FCLT: $\quad \sqrt{n}[\frac{1}{n} S(nt) - pt] \stackrel{d}{\Longrightarrow} BM(0, pq)$, $\quad$ since Var $\Delta_1 = pq$.

Approx: $\quad S \approx BM(p, pq)$.

Bernoulli $\Rightarrow$ Poisson

$$\forall n,\ iid \qquad \Delta_k^n = \begin{cases} 1 & \quad p_n = \frac{\lambda}{n} \\ & \text{wp} \\ 0 & \quad q_n = 1 - p_n \end{cases} \qquad , \quad S^n(t) = \sum_{k=1}^{\lfloor t \rfloor} \Delta_k^n\ , \quad t \geq 0\,.$$

Rare Events: $S^n(nt) \sim$ Binomial $(\lfloor nt \rfloor,\ \frac{\lambda}{n}) \Longrightarrow$ Poisson $(\lambda t),\ t \geq 0$.



$S^n(t),\ 0 \leq t \leq 10n;\ \lambda = 1.$

$S^n(t),\ 0 \leq t \leq n;\ \lambda = 50.$

Equivalently,

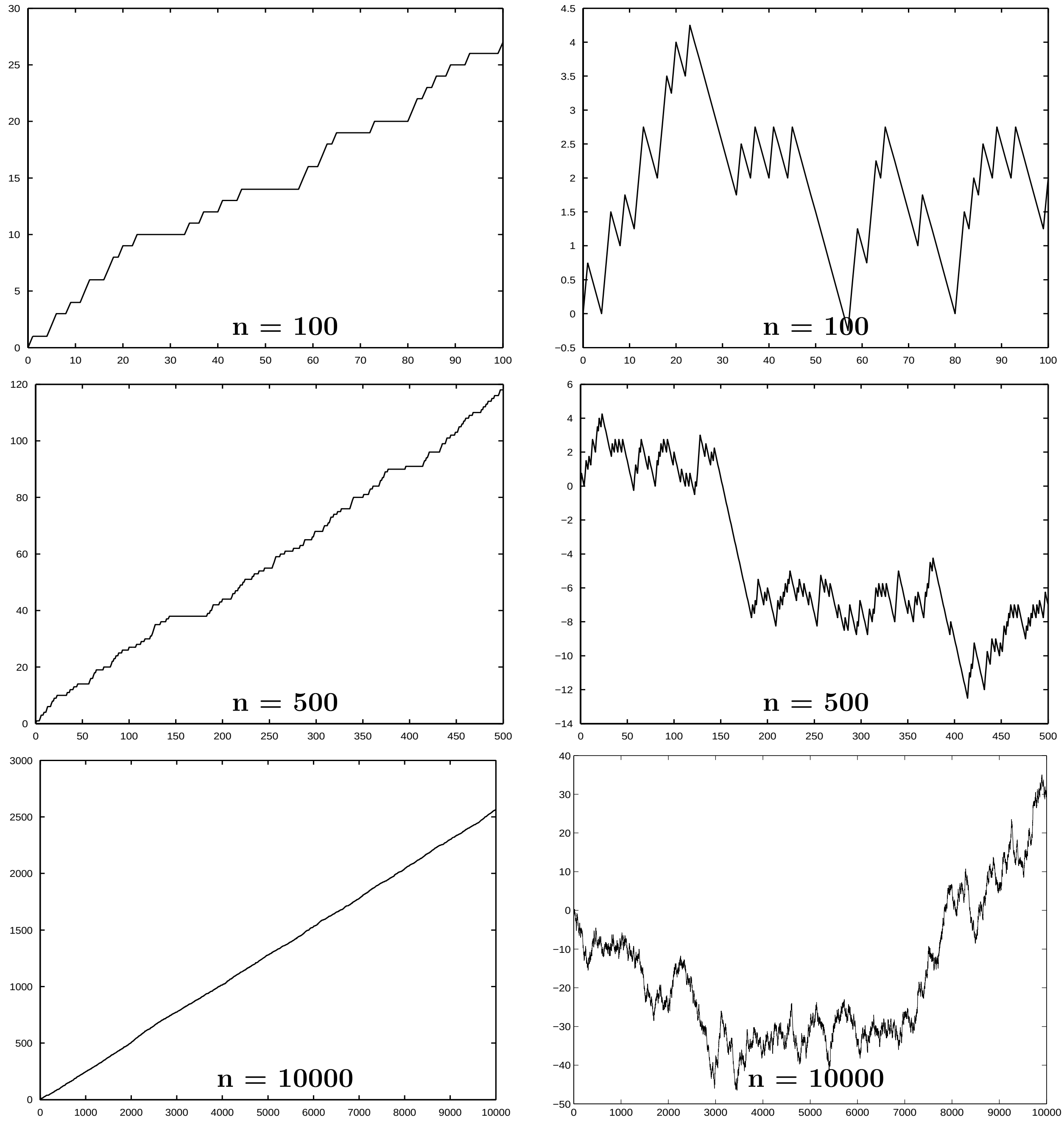
$S^n(nt),\ 0 \leq t \leq 10;\ \lambda = 1;$

But then change scale of X-axis.

Bernoulli $\Rightarrow$ **Brownian** (Review)

$$iid \qquad \Delta_k = \begin{cases} 1 & \\ & \text{wp} \\ 0 & \end{cases} \begin{matrix} p = 0.25 \\ \\ q \end{matrix} \qquad , \quad S(t) = \sum_{k=1}^{\lfloor t \rfloor} \Delta_k^n \,, \quad t \geq 0\,.$$

FSLLN: $\frac{1}{n} S(nt) \xrightarrow{wp1} p \cdot t$; FCLT: $\sqrt{n}[\frac{1}{n} S(nt) - p \cdot t] \stackrel{d}{\Rightarrow} BM(0, pq)$



$S(t),\ 0 \leq t \leq n$ $S(t) - pt,\ 0 \leq t \leq n$

Appendix: Matlab Code for Generating Realizations

Drawing steps:

```
RW_Steps = floor( rand(1, MAX_STEPS) +
                  ones(1, MAX_STEPS) * STEP_PROB) * STEP_SIZE
```

Where: MAX_STEPS - Maximal required steps.
STEP_PROB - p .
STEP_SIZE - Jumps size.

Plotting $S(n)$:

```
plot(0:N, cumsum([0 RW_Steps(1:N)]))
```

Where: N - Number of steps.

Plotting $S(n) - n\mu$:

```
plot(0:N, cumsum([0 RW_Steps(1:N) - ones(1, N) * STEP_SIZE * STEP_PROB]))
```

Where: N - Number of steps.