

Sections 4.1,4.2

Law of Large Numbers (LLN), and Central Limit Theorem (CLT)
(Binomial version)

Ex) Simple experiment with repeated trials

Flip a biased coin, which is heads with probability p , n times.

Let $X_i \sim \text{Bern}(p)$ count a success if the i^{th} flip is heads so

$$S_n = X_1 + X_2 + \dots + X_n \sim \underline{\text{Bin}(n, p)}$$

Counts the total number of heads.

$$\text{Pmf: } P(S_n = k) = \underline{\binom{n}{k} p^k (1-p)^{n-k}}$$

But actually calculating the values of pmf becomes cumbersome when n is large (Matlab fails when $n \approx 2000$).

Binomial approximations (rough version)

Fortunately, when n is large, S_n has simple (easy to calculate) approximations.

First order approximation (S_n is roughly constant):

$$S_n \approx \mathbb{E} S_n = np$$

Second order approximation (Deviation from mean is Gaussian):

$$\begin{aligned} S_n &\approx \mathbb{E} S_n + \sqrt{\text{Var}(S_n)} \cdot N(0,1) \\ &= np + \sqrt{np(1-p)} \cdot N(0,1) \end{aligned}$$

Rigorous versions

Constant approximation: Law of large numbers (LLN)

Gaussian approximation: Central limit theorem (CLT)

Theorem (Law of Large Numbers (LLN)):

$$\frac{S_n}{n} \approx p$$

For any $\epsilon > 0$, $\lim_{n \rightarrow \infty} P\left(\left|\frac{S_n}{n} - p\right| < \epsilon\right) = 1$, $\lim_{n \rightarrow \infty} P\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) = 0$

- E.g, let $p = \frac{1}{2}$, pick $\epsilon = 10^{-10}$. LLN says:

$\lim_{n \rightarrow \infty} P\left(\left|\frac{S_n}{n} - \frac{1}{2}\right| < 10^{-10}\right) = 1$, i.e., for any $\delta > 0$, there exists $N > 0$ so that for $n > N$, $P\left(\left|\frac{S_n}{n} - \frac{1}{2}\right| < 10^{-10}\right) > 1 - \delta$

- “ $\frac{S_n}{n}$ converges in probability to p ”

- $\frac{S_n}{n}$ is the proportion of successes. It has expected value p

Ex) Roll a die n times. Let S_n be the number of sixes rolled. Show that for n large enough, $P\left(S_n \leq \frac{n}{6} \cdot (1.01)\right) \geq 0.99$.

$$S_n \sim \text{Bin}(n, 1/6)$$

$$> \mathbb{E}S_n$$

$$\mathbb{E}S_n = n \cdot 1/6$$

$$P(S_n \leq \frac{n}{6} (1.01)) = P\left(\frac{S_n}{n} \leq \frac{1}{6} + \frac{0.01}{6}\right) = P\left(\frac{S_n}{n} - \frac{1}{6} \leq \frac{0.01}{6}\right)$$

$$\geq P\left(\left|\frac{S_n}{n} - \frac{1}{6}\right| \leq \frac{0.01}{6}\right)$$

$$\geq P\left(\left|\frac{S_n}{n} - \frac{1}{6}\right| < \frac{0.01}{6}\right) = (*)$$

$$\xrightarrow{n \rightarrow \infty} 1 \quad \text{by LLN}$$

$$\text{Fact: } P(X < a) \geq P(|X| < a)$$

Thus, for n large enough, $(*) \geq 0.99$

Ex) You go to the casino with $\overbrace{1}^{10^{12}}$ trillion dollars. You repeatedly play a game which costs 1\$ to play. If you win, you get 100\$. The chance of winning is $1/101$. What is

$$\lim_{n \rightarrow \infty} P(\underbrace{\text{You go broke after playing } \leq n \text{ times}}_{A_n})?$$

$Y_n =$ net gain/loss after playing n times, assuming a
bank roll

$$\text{Note: } P(A_n) \geq P(Y_n \leq -10^{12})$$

$$Y_n = 100 \cdot S_n - n$$

$$S_n \sim \underbrace{\text{Bin}(n, \frac{1}{101})}_{\text{\# of wins}}$$

$$\begin{aligned}
P(X_n \leq -10^{12}) &= P(100 S_n - n \leq -10^{12}) = P\left(\frac{S_n}{n} - \frac{1}{100} \leq -\frac{10^{10}}{n}\right) \\
&= P\left(\frac{S_n}{n} - \frac{1}{101} \leq \frac{1}{100} - \frac{1}{101} - \frac{10^{10}}{n}\right) \\
&\geq P\left(\left|\frac{S_n}{n} - \frac{1}{101}\right| \leq \underbrace{\frac{1}{100} - \frac{1}{101}}_{2\epsilon := \frac{1}{100} - \frac{1}{101} > 0} - \frac{10^{10}}{n}\right) = (*)
\end{aligned}$$

For n large enough $\frac{10^{10}}{n} \leq \epsilon$. Then

$$(*) \geq P\left(\left|\frac{S_n}{n} - \frac{1}{101}\right| \leq \underbrace{2\epsilon}_{\epsilon} - \epsilon\right) \xrightarrow{n \rightarrow \infty} 1 \quad \text{by LLN}$$

$$\Rightarrow \boxed{P(\text{going broke}) \xrightarrow{n \rightarrow \infty} 1}$$

Now, Gaussian approximation

Recall: $S_n \sim \text{Bin}(n, p)$, $\mathbb{E}S_n = np$, $\text{Var}(S_n) = np(1 - p)$.

Standardize: $Q_n := \frac{S_n - np}{\sqrt{np(1-p)}}$ so $E Q_n = \underline{0}$, $\text{Var}(Q_n) = \underline{1}$.

Theorem (Central Limit Theorem (CLT)): $Q_n \rightarrow N(0,1)$ in distribution.
This means for any interval $[a, b]$,

$$\lim_{n \rightarrow \infty} P(Q_n \in [a, b]) = P(\underline{N(0,1)} \in [a, b])$$

- $P(N(0,1) \in [a, b]) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \Phi(b) - \Phi(a)$
- a may be $-\infty$, b may be ∞

Ex) $S_n \sim \text{bin}\left(n, \frac{1}{2}\right)$. What is $\lim_{n \rightarrow \infty} P\left(S_n \geq \frac{n}{2}\right)$?

$$\mathbb{E} S_n = n/2$$

$$\text{var}(S_n) = n/4$$

$$P(S_n \geq \frac{n}{2}) = P(S_n - \frac{n}{2} \geq 0) = P(\overset{Q_n}{\frac{S_n - n/2}{\sqrt{n/4}}} \geq 0)$$

$$\xrightarrow{n \rightarrow \infty} P(N(0,1) \geq 0)$$

by CLT

$$= \boxed{1/2}$$

Normal approximation:

CLT says $\frac{S_n - np}{\sqrt{np(1-p)}} \rightarrow N(0,1)$.

Roughly this means for large n , $\frac{S_n - np}{\sqrt{np(1-p)}}$ is approximately distributed as $N(0,1)$.

Thus, S_n is approx. $N(0,1) \cdot \sqrt{np(1-p)} + np = \underline{N(np, np(1-p))}$.

Handwritten notes: ES_n with a green arrow pointing down to np ; $Var(S_n)$ with a green squiggly line above $np(1-p)$.

Normal approximation: Replace S_n with $N(np, np(1-p))$ when calculating probabilities.

In other words, replace S_n with a normal r.v. that has the same mean and variance as S_n .

Rule of thumb: Approximation is accurate if $np(1 - p) > 10$.

Ex) Flip a fair coin 10,000 times. Approximate the probability that you get between 4850 and 5100 heads. Let S_n be number of heads.

$$S_n \sim \underline{B_{10,000}(\frac{1}{2})} \approx N(\underline{5000}, \underline{\frac{10,000}{4}})$$

$$\begin{aligned} P(4850 \leq S_n \leq 5100) &\approx \underline{P(4850 \leq N(5000, 2500) \leq 5100)} \\ &= P\left(\frac{4850 - 5000}{\sqrt{2500}} \leq \frac{N(5000, 2500) - 5000}{\sqrt{2500}} \leq \frac{5100 - 5000}{\sqrt{2500}}\right) \\ &= P(-3 \leq N(0,1) \leq 2) = \boxed{\Phi(2) - \Phi(-3)} \end{aligned}$$

Ex) Roll a die 6,000 times. Approximate the probability of at least 1100 sixes.

$$S_n = \# \text{ of sixes} \sim \text{Bin}(6,000, \frac{1}{6}) \approx N(1000, \underbrace{6000 \cdot \frac{1}{6} \cdot \frac{5}{6}}_{\text{Var}(S_n)})$$

$\uparrow$
 $E S_n$

normal approx

$$= N(1000, \frac{5000}{6})$$



$$P(S_n \geq 1100) \approx P(N(1000, \frac{5000}{6}) \geq 1100)$$

$$= P(1000 + \sqrt{\frac{5000}{6}} N(0,1) \geq 1100)$$

$$= P(N(0,1) \geq \frac{100}{\sqrt{\frac{5000}{6}}}) = \boxed{1 - \Phi\left(\frac{100}{\sqrt{\frac{5000}{6}}}\right)}$$

Random walk

Ex) Each day, take a step left or right with equal probability. About how far are you from home after n steps?

$$\text{Let } X_i = \begin{cases} 1 & \text{w.p. } 1/2 \\ -1 & \text{w.p. } 1/2 \end{cases} \quad i=1, \dots, n$$

$$\text{Then position after } n \text{ steps} = S_n = \sum_{i=1}^n X_i$$

$$\text{Normal approx: } S_n \approx N(\mathbb{E}S_n, \text{Var}(S_n))$$

$$\mathbb{E}S_n = n \cdot \mathbb{E}X_1 = 0$$

$$\text{Var}(S_n) = n \cdot \text{Var}(X_1) = n (\cancel{\mathbb{E} X_1^2} - (\cancel{\mathbb{E} X_1})^2) = n$$

$$\text{So, } S_n \approx N(0, n) = \sqrt{n} N(0, 1).$$

Thus, after n steps, distance from home is on the order of $\sqrt{n}$.

Concentration of normal r.v.

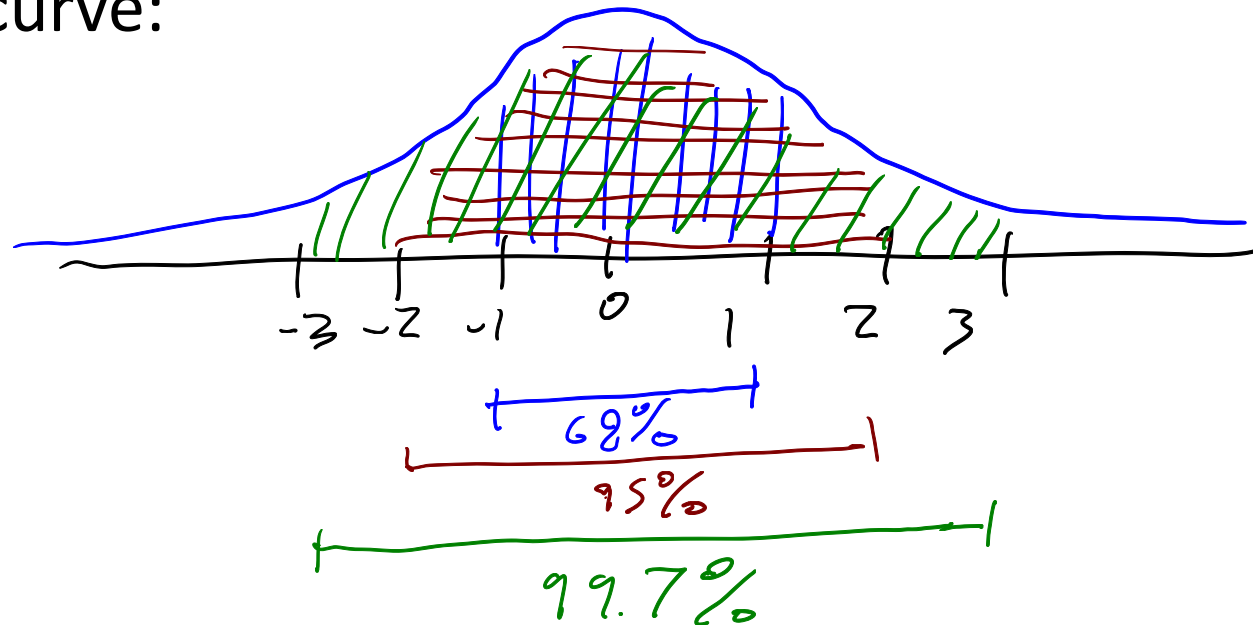
Fact: A $N(0,1)$ r.v. is very unlikely to deviate far from 0.

$$P(-1 \leq N(0,1) \leq 1) \approx 0.68$$

$$P(-2 \leq N(0,1) \leq 2) \approx 0.95$$

$$P(-3 \leq N(0,1) \leq 3) \approx 0.997$$

Area under curve:



Concentration for general normal r.v.

Implication: With probability ≈ 0.997 .

$N(\mu, \sigma^2) \in [\mu - 3\sigma, \mu + 3\sigma]$.

W.P. ≈ 0.95

$N(\mu, \sigma^2) \in [\mu - 2\sigma, \mu + 2\sigma]$

- “Normal random variable is almost always within three standard deviations of the mean.”
- Combined with normal approximation, we have

$$\text{Bin}(n, p) \in [\underline{np - 3\sqrt{np(1-p)}}, \overline{np + 3\sqrt{np(1-p)}}]$$

with probability approaching 0.997.