MATH 180A: INTRO TO PROBABILITY
(FOR DATA SCIENCE)

www.math.ucsd.edu/~tkemp/180A

Today: § 4.1 – 4.3

Next: § 4.3 – 4.4

Midterm grades released; regrade requests Wed 10/30 8am–11pm

Lab 3 grades released; regrade requests Thu 10/31 8am–11pm

HW 4 due Friday by 11:59pm

Lab 4 due next Wednesday (Nov 6) by 11:59pm

Poisson Approximation of Binomial

Fix $\lambda > 0$. For each n, let $T_n \sim \mathrm{Bin}(n, \lambda/n)$.
For any fixed $k \in \mathbb{N}$,

$$\lim_{n\to\infty} \mathbb{P}(T_n = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

In particular, for $a \le b$ positive numbers,

$$\lim_{n\to\infty} \mathbb{P}(a \le T_n \le b) = \sum_{\substack{a \le k \le b \\ k \in \mathbb{N}}} e^{-\lambda} \frac{\lambda^k}{k!}.$$

Normal Approximation of Binomial

Fix $p \in (0,1)$. For each n, let $S_n \sim \mathrm{Bin}(n,p)$.
For any fixed $a \le b$,

$$\lim_{n\to\infty} \mathbb{P}\left(a \le \frac{S_n - np}{\sqrt{np(1-p)}} \le b \right) = \underbrace{\int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx}_{\Phi(b) - \Phi(a)}.$$

Question:

A fair 20-sided die is tossed 400 times. We want to calculate the probability that a 13 came up at least 25 times. We should use:

$T \sim \mathrm{Bin}(400, \frac{1}{20})$

$\mathrm{Bin}(n, \lambda/n)$

(a) Poisson Approximation.
(b) Normal Approximation.
(c) Either.
(d) Neither.

$T \sim \mathrm{Bin}(400, \frac{1}{20})$

$\mathbb{P}(T \geq 25) = 1 - \mathbb{P}(T \leq 24)$

$= 1 - \sum_{k=0}^{24} \binom{400}{k} \left(\frac{1}{20}\right)^k \left(\frac{19}{20}\right)^{400-k}$

$= \boxed{15.10\%}$

12.57%

$\|$

$1 - \left(\Phi\left(\frac{5}{\sqrt{19}}\right) - 0\right)$

$\|$

$1 - \int_{-\infty}^{5/\sqrt{19}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx$

Poisson: $n = 400, \quad \lambda = 20$

$\mathbb{P}(T \geq 25) = 1 - \mathbb{P}(T < 25)$

$= 1 - \sum_{k=0}^{24} \mathbb{P}(T = k)$

$\approx 1 - \sum_{k=0}^{24} e^{-20} \frac{(20)^k}{k!}$

$= \boxed{15.68\%}$

Normal Approx

$\mathbb{E}(T) = 400 \cdot \frac{1}{20} = 20$

$\mathrm{Var}(T) = 400 \cdot \frac{1}{20} \cdot \frac{19}{20} = 19$

$0 \leq$

$\mathbb{P}(T \geq 25) = 1 - \mathbb{P}(T < 25)$

$= 1 - \mathbb{P}(T - 20 < 5)$

$\approx \quad = 1 - \mathbb{P}\left(\frac{T-20}{\sqrt{19}} < \frac{5}{\sqrt{19}}\right)$

Eg. A fair die is rolled 720 times. What is the probability that exactly 113 sixes come up?

$S = \#\text{sixes} \sim \text{Bin}(720, \frac{1}{6})$. $\mathbb{P}(S=113) = \binom{720}{113}\left(\frac{1}{6}\right)^{113}\left(\frac{5}{6}\right)^{607} \doteq 3.184\%$

Normal Approximation:

$\mathbb{E}(S) = 720 \cdot \frac{1}{6} = 120 \qquad Var(S) = 720 \cdot \frac{1}{6} \cdot \frac{5}{6} = 100$

$$\mathbb{P}(S=113) = \mathbb{P}(113 \le S \le 113) = \mathbb{P}\left(-0.7 \le \frac{S-120}{\sqrt{100}} \le -0.7\right) \approx \Phi(-0.7) - \Phi(-0.7) = 0.$$

$$= \mathbb{P}(112.5 \le S \le 113.5) = \mathbb{P}\left(-0.75 \le \frac{S-120}{10} \le -0.65\right)$$

$$= \Phi(-0.65) - \Phi(-0.75)$$

$$\doteq 3.122\%$$

"Continuity correction!"

$$\mathbb{P}(k_1 \le S \le k_2) = \mathbb{P}\left(k_1 - \tfrac{1}{2} \le S \le k_2 + \tfrac{1}{2}\right) \approx \Phi\left(\frac{k_2 + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\left(k_1 - \tfrac{1}{2} - np\right)/\sqrt{np(1-p)}\right)$$

4.2

What is $\mathbb{E}$?

Precise definition: $\mathbb{E}(X) = \sum_k k \cdot \mathbb{P}(X=k)$

Intuition: Sample X independently many times, compute the average!

Binomial $S_n \sim \mathrm{Bin}(n,p)$ $\quad \mathbb{E}(S_n) = np$

$\quad \overset{\shortparallel}{X_1 + \cdots + X_n}$, $\quad X_j \sim \mathrm{Ber}(p)$ $\quad \mathbb{E}(X_j) = p$. $\quad \mathbb{E}\left(\frac{S_n}{n}\right) = \frac{np}{n} = \boxed{p}$

<u>Theorem</u> (Law of Large Numbers for Bernoulli Trials)

Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli trials with success probability p. Then " $\underbrace{\frac{X_1 + \cdots + X_n}{n}}_{S_n/n} \to p$ as $n \to \infty$"

Precisely: For any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - p\right| < \varepsilon\right) = 1$$

If $S_n \sim \mathrm{Bin}(n,p)$ (p fixed), and $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - p\right| > \varepsilon\right) = 0.$$

Proof:

$$\frac{S_n}{n} - p = \frac{S_n - np}{n}\,\frac{\sqrt{np(1-p)}}{\sqrt{np(1-p)}} = \underbrace{\frac{S_n - np}{\sqrt{np(1-p)}}}_{\to\, N(0,1)} \cdot \underbrace{\frac{\sqrt{n(p)(1-p)}}{n}}_{\frac{\sqrt{p(1-p)}}{\sqrt{n}}}$$

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| < \varepsilon\right) = \mathbb{P}\left(\left|\frac{S_n - np}{\sqrt{np(1-p)}}\right| < \underbrace{\frac{\varepsilon}{\sqrt{p(1-p)}}\cdot\sqrt{n}}_{\text{large \#}}\right)$$

Pick your favorite large R. Find some n_0 s.t.

$$\frac{\varepsilon}{\sqrt{p(1-p)}}\sqrt{n_0} > R.$$

$$\approx \underbrace{\mathbb{P}(|X| < R)}_{X \sim N(0,1)} \to 1 \text{ as } R \to \infty.$$

///

Example

Flip a fair coin n times. How does

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\#\text{Heads}}{n} \geq 50.01\%\right) = 0.$$

behave as $n \to \infty$?

Suppose after $10{,}000$ flips, there are $5{,}001$ Heads.
Should we doubt that the coin is really fair?

What if, after $1{,}000{,}000$ flips, there are $500{,}100$ Heads.
Now how confident should we be that the coin is really fair?

$$\mathbb{P}\left(\frac{S_n}{n} \geq \frac{1}{2} + \varepsilon\right) = \mathbb{P}\left(\frac{S_n - \frac{1}{2}n}{\sqrt{n\cdot\frac{1}{2}\cdot\frac{1}{2}}} \geq 2\varepsilon\sqrt{n}\right) \approx \mathbb{P}\left(X \geq 2\varepsilon\sqrt{n}\right)$$

$\varepsilon = 0.01$

$X \sim N(0,1)$