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

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

Today: § 4.4 – 4.5

Next: § 4.6

Lab4 due Wednesday (Nov 6) by 11:59pm

HW5 due Friday (Nov 8) by 11:59pm

Poisson vs. Normal Approximation – Quantitative

4.4

Theorem. Let $S_n \sim \mathrm{Bin}(n,p)$

$X \sim \mathrm{Poisson}(np)$

$Y \sim \mathcal{N}(0,1)$

For any subset $A \subseteq \mathbb{N}$,

$$|\mathbb{P}(S_n \in A) - \mathbb{P}(X \in A)| \le np^2$$

if $p = \frac{\lambda}{n^{0.51}}$ $(\lambda > 0)$

$np^2 = n\left(\frac{\lambda}{n^{0.51}}\right)^2 = \frac{\lambda^2 n}{n^{1.02}}$

$\to 0$ as $n \to \infty$

OTOH, for any $x \in \mathbb{R}$,

Berry–Esseen Thm.

$$\Bigg|\underbrace{\mathbb{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \le x\right)}_{\text{CDF of } \frac{S_n - np}{\sqrt{np(1-p)}}} - \underbrace{\mathbb{P}(Y \le x)}_{\Phi(x)}\Bigg| \le \frac{3}{\sqrt{np(1-p)}}$$

3 is not optimal

optimal

Upshot: if np^2 is small, use Poisson Approximation.

if $np(1-p)$ is quite large, use Normal Approximation.

Beyond independent trials:

* The normal approximation breaks down qickly if the trials are dependent.
* The Poisson approximation holds up well under "weak dependence"

Example. A factory experiences 3 accidents per month, on average. What is the probability there will be 3 accidents this month?

X = # accidents in a given month. well modeled by a Poisson.

$X \sim \text{Poisson}(\lambda)$

$3 = \mathbb{E}(X) = \lambda$

$$\mathbb{P}(X=3) = e^{-3}\frac{3^3}{3!} \doteq 22.4\%$$

$$\frac{3^3}{3!} = \frac{3^2\cdot\cancel{3}}{\cancel{3}\cdot 2\cdot 1} = \frac{3^2}{2!}.$$

$$\mathbb{P}(X=2) = e^{-3}\frac{3^2}{2!} \doteq 22.4\%$$

Wait Times

4.5

Question: You're tossing a fair die until you get a 6. It's been 12 tosses already, but no 6's so far. The time you have to wait until the first 6 from now is

(a) Less than
(b) The same as
(c) Greater than

the time you would wait from the start.

"Gambler's Fallacy"

Time of first success $T \sim \text{Geom}(p)$.

$$\mathbb{P}(T > t) = \sum_{k=t+1}^{\infty} (1-p)^{k-1} p = p(1-p)^t \sum_{\ell=0}^{\infty} (1-p)^\ell = \cancel{p} \cdot (1-p)^t \cdot \frac{1}{\underbrace{1-(1-p)}_{1/\cancel{p}}}$$

$$F_T(t) = \mathbb{P}(T \le t) = 1 - (1-p)^t$$

Given that we've waited longer than t, what is the probability that we'll have to wait more than s more?

$$\mathbb{P}(T > t+s \mid T > t)$$

$$= \frac{\mathbb{P}(T > t+s \;\&\; T > t)}{\mathbb{P}(T > t)} \leftarrow = \frac{\mathbb{P}(T > t+s)}{\mathbb{P}(T > t)} = \frac{(1-p)^{t+s}}{(1-p)^t}$$

$$= \frac{\cancel{(1-p)^t}(1-p)^s}{\cancel{(1-p)^t}}$$

$$= (1-p)^s$$

$$= \mathbb{P}(T > s)$$

"memoryless!"

Continuous Wait Times

In the real world, most wait times are continuous random variables.

Eg. After a lull, the arrival time of the first customer at a post office.

Eg. The time until a radioactive particle decays.

These wait times are continuous, but have the same defining memoryless property:

$G(t) = 1 - F_T(t)$

$$\mathbb{P}(T > t+s) = \mathbb{P}(T > s)\,\mathbb{P}(T > t)$$

Set $\mathbb{P}(T > t) = G(t)$. Thus $G(t+s) = G(t)G(s)$ $\quad (*)$

Theorem: If $G : \mathbb{R}_+ \to \mathbb{R}_+$ is a cont. differentiable function satisfying $(*)$, then for some $a \in \mathbb{R}$,

$$G(t) = e^{at}$$

$$e^{a(t+s)} = e^{at} e^{as} \checkmark$$

Proof: $\frac{\partial}{\partial s} G(t+s) = \frac{\partial}{\partial s}[G(t)G(s)]$

$$\frac{d}{ds}(t+s) \cdot G'(t+s) = G(t)G'(s) \qquad \text{Let } s \downarrow 0.$$

$$G'(t) = \underbrace{G'(0)}_{a} G(t) \qquad G'(t) = aG(t)$$

$$\Rightarrow G(t) = Ce^{at}$$

$$\therefore \underline{c^2 = c} \qquad c = 0, \boxed{1.}$$

$$Ce^{a(t+s)} = G(t+s) = G(t)G(s) = Ce^{at} \cdot Ce^{as} = C^2 e^{a(t+s)}$$

Now, if $G(t) = \mathbb{P}(T > t) = 1 - \mathbb{P}(T \leq t)$, we get

$$F_T(t) = \mathbb{P}(T \leq t) = 1 - G(t) = 1 - e^{at} \to 1 \text{ as } t \to \infty$$

$$\therefore a < 0 \qquad \text{Let } \lambda = -a$$

Definition: The **exponential distribution** with parameter $\lambda > 0$ is given by CDF

$$F(t) = \begin{cases} 1 - e^{-\lambda t}, & t \geq 0 \\ 0, & t < 0 \end{cases} \qquad f(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

Mean and Variance of $T \sim \mathrm{Exp}(\lambda)$

$$\mathbb{E}(T) = \int_0^\infty t\, f_T(t) = \int_0^\infty t \cdot \lambda e^{-\lambda t}\, dt = \text{CALCULUS} = \frac{1}{\lambda}.$$

$$\mathrm{Var}(T) = \frac{1}{\lambda^2} \quad \text{(similar calculation)}$$

E.g. The average phone call is 5 minutes in length. What is the probability your next phone call will be longer than 3 minutes?

$$T \sim \mathrm{Exp}(\lambda), \quad \frac{1}{\lambda} = \mathbb{E}(T) = 5$$

$$\therefore \mathbb{P}(T > 3) = \int_3^\infty \underbrace{\frac{1}{5} e^{-\frac{1}{5}t}} dt = \left(-e^{-\frac{t}{5}}\right)\Big|_{t=3}^{t=\infty} = e^{-\frac{3}{5}} \approx 54{,}9\%$$

$$\mathbb{P}(T > x) = e^{-x/5} = e^{-\lambda x}$$

E.g. On a forest road, cars come by Turtle Rock on average every 30 minutes. Tianyi the Turtle needs 10 minutes to cross the road. What is the probability she can cross safely?

T = arrival time of next car, $T \sim \text{Exp}(\frac{1}{30})$

$$\mathbb{P}(T > 10) = e^{-\frac{1}{30} \cdot 10} = e^{-\frac{1}{3}} \doteq 71.7\%.$$

Just before she starts to cross, Li-Tien the lemur tells her he's been hanging around for over 20 minutes and no cars have come by. Does this change Tianyi's mind about how safe it is to cross?

$$\mathbb{P}(T > 20+10 \mid T > 20) = \mathbb{P}(T > 10) = 71.7\%$$