LECTURE 20

- **Readings:** Section 5.4

Lecture outline

- Review
- Central limit theorem
- Normal approximations
- De Moivre–Laplace binomial approximation

Review: Convergence in probability

- Sequence of random variables $Y_1, Y_2, \ldots$ **converges in probability to a number** a when

$$\text{for any } \epsilon > 0, \quad \lim_{n\to\infty} \mathbf{P}(|Y_n - a| \geq \epsilon) \;=\; 0$$

- Examples of $Y_1, Y_2, \ldots$ converging in probability to 0:
  - Y_n continuous uniform over $[0, 1/n]$
  - Y_n continuous uniform over $[-1/n, 1/n]$
  - $p_{Y_n}(y) \;=\; \begin{cases} 1-\frac{1}{n}, & y = 0 \\ \frac{1}{n}, & y = 1 \end{cases}$
  - $p_{Y_n}(y) \;=\; \begin{cases} 1-\frac{1}{n}, & y = 0 \\ \frac{1}{n}, & y = n^2 \end{cases}$
  - $p_{Y_n}(y) \;=\; \begin{cases} 1-\frac{1}{n}, & y = e^{-n} \\ \frac{1}{n}, & y = e^{n} \end{cases}$

Central limit theorem

- Let Z be a standard normal r.v. (zero mean, unit variance), $X_1, X_2, \ldots$ i.i.d. with finite mean μ and variance σ^2, and

$$Z_n \;=\; \frac{(X_1 + X_2 + \cdots + X_n) - n\mu}{\sqrt{n}}.$$

$$\mathbf{E}[Z_n] \;=\; 0 \qquad\qquad \text{var}(Z_n) \;=\; 1$$

For every z,

$$\lim_{n\to\infty} \mathbf{P}(Z_n \leq z) \;=\; \mathbf{P}(Z \leq z) \;=\; \Phi(z).$$

- Asymptotic equality of CDFs
- Not a statement about PDFs or CDFs
- Standard normal CDF $\Phi(z)$ is built in to many packages and available in tables
- Proof is not too difficult with transforms

Normal approximation

- Treat Z_n as if standard normal
  - Makes S_n have the $\mathcal{N}(n\mu, n\sigma^2)$ distribution
- How good is it when n is "moderate"?
  - Never exactly correct (unless X_is are normal)
  - Speed of convergence depends on X_i distribution
  - Convergence is faster with symmetric distributions
  - Convergence is faster with unimodal distributions

The pollster's problem revisted

- f: fraction of population that "..."
- ith (randomly selected) person polled: $X_i = \begin{cases} 1, & \text{if yes;} \\ 0, & \text{if no.} \end{cases}$
- $M_n = (X_1 + \cdots + X_n)/n$ is fraction of "yes" in our sample
- Goal: "95% confidence in being within 1% error"

$$\mathbf{P}(|M_n - f| \geq 0.01) \leq 0.05$$

- Use normal approximation: treat M_n as normal

The pollster's problem revisted (2)

- Event of interest (to be made unlikely): $|M_n - f| \;\geq\; 0.01$

$$\left|\frac{X_1 + \cdots + X_n - nf}{n}\right| \;\geq\; 0.01$$

$$\left|\frac{X_1 + \cdots + X_n - nf}{\sqrt{n}\sigma}\right| \;\geq\; \frac{0.01\sqrt{n}}{\sigma}$$

$$\begin{aligned} \mathbf{P}(|M_n - f| \geq 0.01) \;&\approx\; \mathbf{P}(|Z| \geq 0.01\sqrt{n}/\sigma) \\ &\leq\; \mathbf{P}(|Z| \geq 0.02\sqrt{n}) \\ &=\; 2\,\mathbf{P}(Z < -0.02\sqrt{n}) \;=\; 2\left(1 - \Phi(0.02\sqrt{n})\right) \end{aligned}$$

- Pick n large enough that $\Phi(0.02\sqrt{n}) \;\geq\; 1 - \dfrac{0.05}{2} \;=\; 0.975$
  - $\Phi(1.96) \approx 0.975$, so $n = 9604$ suffices

The pollster's problem revisted (3)

- Why is Chebyshev-inspired pollster so conservative?
- Fix $p \in (0,1)$ and let $X_1, X_2, \ldots$ be independent Bernoulli(p)
- $S_n = X_1 + \cdots + X_n$ is binomial(n, p)
  - mean np, variance $np(1-p)$
- CDF of $\dfrac{S_n - np}{\sqrt{np(1-p)}}$ approaches standard normal CDF

Approximating a binomial: Example

- $n = 36$, $p = 0.5$; find $\mathbf{P}(S_n \le 21)$
- Exact answer:

$$\mathbf{P}(S_n \le 21) \;=\; \sum_{k=0}^{21} \binom{36}{k} \left(\frac{1}{2}\right)^{36} \;\approx\; 0.8785$$

- Normal approximation:

$$\mathbf{P}(S_n \le 21) \;=\; \mathbf{P}\left(\frac{S_n - 36 \cdot 0.5}{\sqrt{36 \cdot 0.5 \cdot 05}} \le \frac{21 - 36 \cdot 0.5}{\sqrt{36 \cdot 0.5 \cdot 05}}\right) \;\approx\; \Phi(1) \;\approx\; 0.8413$$

- But also:

$$\begin{aligned}\mathbf{P}(S_n \le 21) \;&=\; \mathbf{P}(S_n < 22) \\ &=\; \mathbf{P}\left(\frac{S_n - 36 \cdot 0.5}{\sqrt{36 \cdot 0.5 \cdot 05}} < \frac{22 - 36 \cdot 0.5}{\sqrt{36 \cdot 0.5 \cdot 05}}\right) \;\approx\; \Phi\left(\frac{4}{3}\right) \;\approx\; 0.9088\end{aligned}$$

De Moivre–Laplace approximation

- Using the midpoint of discrete values is called the De Moivre–Laplace approximation:

$$\begin{aligned}\mathbf{P}(S_n \le 21) \;&=\; \mathbf{P}(S_n \le 21.5) \\ &=\; \mathbf{P}\left(\frac{S_n - 36 \cdot 0.5}{\sqrt{36 \cdot 0.5 \cdot 05}} \le \frac{21.5 - 36 \cdot 0.5}{\sqrt{36 \cdot 0.5 \cdot 05}}\right) \;\approx\; \Phi\left(\frac{7}{6}\right) \;\approx\; 0.8783\end{aligned}$$

De Moivre–Laplace approximation for binomial PMF

- "1/2 correction" enables a reasonable PMF approximation:

$$\mathbf{P}(S_n = 19) \;=\; \mathbf{P}(18.5 \le S_n \le 19.5)$$

$$18.5 \le S_n \le 19.5 \iff \frac{18.5 - 18}{3} \le \frac{S_n - 18}{3} \le \frac{19.5 - 18}{3}$$

$$\iff \frac{1}{6} \le Z_n \le \frac{1}{2}$$

$$\begin{aligned}\mathbf{P}(S_n = 19) \;&\approx\; \mathbf{P}\left(\frac{1}{6} \le Z \le \frac{1}{2}\right) \\ &=\; \Phi\left(\frac{1}{2}\right) - \Phi\left(\frac{1}{6}\right) \;\approx\; 0.6915 - 0.5675 \;=\; 0.124\end{aligned}$$

- Exact answer: $\binom{36}{19}\left(\frac{1}{2}\right)^{36} \;\approx\; 0.1251$

Poisson vs. normal approximations of the binomial

- Poisson arrivals during unit interval equals: sum of n (independent) Poisson arrivals during n intervals of length $1/n$
  - Let $n \to \infty$, apply CLT (?)
  - Poisson $\stackrel{?}{=}$ normal
- Binomial(n, p)
  - p fixed, $n \to \infty$: normal
  - np fixed, $n \to \infty$, $p \to 0$: Poisson
- Single instance: use Poisson when p is very small
  - $p = 1/100$, $n = 100$: use Poisson approximation
  - $p = 1/10$, $n = 500$: use normal approximation

Application of the CLT

- Easy to apply: Uses only means and variances
- Useful computational shortcut, even if the distribution of S_n is known
- Justifies noise models involving normal random variables
  - Accumulation of many small, approx.-independent effects
- Normal distribution has many extremal properties
  - maximizing entropy, minimizing Fisher information, Heisenberg uncertainty principle, ...

"Physicists believe that the Gaussian law has been proved in mathematics while mathematicians think that it was experimentally established in physics."
— Henri Poincaré [*Science et Hypothesis*, 1904]