LECTURE 22

- Readings: Section 7.4

Lecture outline

- The Central Limit Theorem:
  - Introduction
  - Formulation and interpretation
  - Pollster’s problem
  - Usefulness

Introduction

- $X_1, \cdots, X_n$ i.i.d. finite variance σ^2
- Look at three variants of their sum:
- $S_n = X_1 + \cdots + X_n$ variance $n\sigma^2$
- $M_n = \frac{S_n}{n}$ variance σ^2/n

  converges "in probability" to $\mathbf{E}[X]$ (WLLN)
- $\frac{S_n}{\sqrt{n}}$ constant variance σ^2

  - Asymptotic shape?

Convergence of the Sample Mean

$X_1, \cdots, X_n$ i.i.d., (finite mean μ and variance σ^2)

$$M_n = \frac{X_1 + \cdots + X_n}{n}$$

- **Mean:** $\mathbf{E}[M_n] = \mu$
- **Variance:** $\mathsf{Var}(M_n) = \frac{\sigma^2}{n}$
- **Chebyshev**: $\mathbf{P}(|M_n - \mathbf{E}[M_n]| \geq \epsilon) \leq \frac{\mathsf{Var}(M_n)}{\epsilon^2}$
- **Limit:** $\mathbf{P}(|M_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$

The Central Limit Theorem

- "Standardized" $S_n = X_1 + \cdots + X_n$:

$$Z_n = \frac{S_n - \mathbf{E}[S_n]}{\sigma_{S_n}} = \frac{S_n - n\mathbf{E}[X]}{\sqrt{n}\sigma}$$

  - zero mean
  - unit variance

- Let Z be a standard normal r.v. (zero mean, unit variance)

- **Theorem**: For every c : $\mathbf{P}(Z_n \leq c) \to \mathbf{P}(Z \leq c)$

- $\mathbf{P}(Z \leq c)$ is the standard normal CDF $\Phi(c)$, available from the normal tables.

What exactly does it say?

- CDF of Z_n converges to normal CDF
  - Not a statement about convergence of PDFs or PMFs.

- **Normal Approximation:**

- Treat Z_n as if normal
  - Also treat S_n as if normal

- **Can we use it when n is "moderate" ?**

- Yes, but no nice theorems in this effect
- Symmetry helps a lot

The Pollster's Problem

- f : fraction of population that do "...........".
- i^{th} person polled: $X_i = \begin{cases} 1 & \text{If "Yes"}. \\ 0 & \text{If "No"}. \end{cases}$
- $M_n = \frac{X_1 + \cdots + X_n}{n}$: fraction of "Yes" in our sample.
- Suppose we want: $\mathbf{P}(|M_n - f| \geq .01) \leq .05$
- Event of interest: $|M_n - f| \geq .01$

$$\left| \frac{X_1 + \cdots + X_n - nf}{n} \right| \geq .01$$

$$\left| \frac{X_1 + \cdots + X_n - nf}{\sqrt{n}\sigma} \right| \geq \frac{.01\sqrt{n}}{\sigma}$$

$$\mathbf{P}(|M_n - f| \geq .01) \approx \mathbf{P}(|Z| \geq 0.01\sqrt{n}/\sigma)$$

$$\leq \mathbf{P}(|Z| \geq 0.02\sqrt{n})$$

The Pollster's Problem

- we want: $\mathbf{P}(|M_n - f| \geq .01) \leq .05$

$$\mathbf{P}(|M_n - f| \geq .01) \approx \mathbf{P}(|Z| \geq 0.02\sqrt{n})$$

$$= 2 - 2\mathbf{P}(Z \leq 0.02\sqrt{n}) \leq .05$$

- From Table: $n \geq 9604$

- Compare to $n \geq 50,000$ that we derived using Chebychev's inequality

Usefulness of the CLT

- Only means and variances matter.

- Much more accurate than Chebyshev’s inequality

- Useful computational shortcut, even if we have a formula for the distribution of S_n

- Justification of models involving normal r.v.’s
  - Noise in electrical components
  - Motion of a particle suspended in a fluid (Brownian motion)