CS 70 ## 4.16 Lecture 24

Sample Means

$M_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$ where all X_i are independent, identically distributed

$E(X_i) = \underline{E(X)}$, the true mean $\sigma_{X_i}^2 = \sigma_X^2$

We run n experiments to estimate the true mean.

$E(M_n) = E\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{E(X_1) + E(X_2) + \ldots + E(X_n)}{n} = \frac{n E(X)}{n} = E(X)$

M_n is an unbiased estimator of $E(X)$.

$\sigma_{M_n}^2 = \text{var}\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{1}{n^2}[\text{var}(X_1) + \text{var}(X_2) + \ldots \text{var}(X_n)] = \frac{n\,\text{var}(X)}{n^2} = \frac{\sigma_X^2}{n}$

Weak Law of Large Numbers

Chebyshev: $P(|M_n - \underbrace{E(M_n)}_{E(X)}| \geq \epsilon) \leq \underbrace{\sigma_{M_n}^2}_{\sigma_X^2/n} / \epsilon^2 \rightarrow P(|M_n - E(X)| \geq \epsilon) \leq \sigma_X^2 / n\epsilon^2$

Then, $\lim_{n \to \infty} \sigma_X^2 / n\epsilon^2 = 0$, so as we increase the number of samples, the probability that it deviates from $E(X)$ by <u>any</u> ϵ approaches 0.

This means the sample mean converges in probability to true mean.

$\rightarrow$ $X_1, X_2, \ldots$ IID RVs w/mean $E(X)$, σ_X^2. For every $\epsilon > 0$, $P(|M_n - E(X)| \geq \epsilon) \to 0$.

Pollster Problem, Confidence Intervals

We want to estimate the true fraction μ of voters that believe marijuana should be legalized. Determine the smallest # of voters we must poll so that we're at least 95% confident that M_n is within $\mu \pm \epsilon$. Say $\epsilon = 0.01$.

μ — $X_i = 1$ legalize $\sigma_X^2 = \mu(1-\mu)$ $\sigma_{M_n}^2 = \frac{\sigma_X^2}{n} = \frac{\mu(1-\mu)}{n}$

$1-\mu$ — $X_i = 0$ don't legalize $E(X) = \mu$ $P_r(|M_n - \mu| < \epsilon) \geq 0.95$

$\frac{\sigma_{M_n}^2}{\epsilon^2} = \frac{\mu(1-\mu)}{n\epsilon^2} \leq 0.05$ $P_r(|M_n - \mu| \geq \epsilon) \leq \frac{\sigma_{M_n}^2}{\epsilon^2} \leq 0.05$

We don't know μ, so let's use the worst case $\mu(1-\mu) \longrightarrow \mu = 0.5$.

$\frac{0.25}{n\epsilon^2} \leq 0.05 \longrightarrow n \geq 0.2/\epsilon^2$

If $\epsilon = 0.01$, you need $n \geq 50{,}000$, which is ridiculous.

Continuous RVs

Uncountably infinite set of outcomes. $P(X = x) = 0$

<u>Cumulative Distribution Function (CDF)</u> $F_X(x) = P(X \leq x)$ $\lim_{x \to \infty} = 1$

$\lim_{x \to -\infty} = 0$

$P(a < X \leq b) = F_X(b) - F_X(a)$