the law of large numbers & the CLT

\[ \Pr \left( \lim_{n \to \infty} \left( \frac{X_1 + \cdots + X_n}{n} \right) = \mu \right) = 1 \]
• Consider i.i.d. (independent, identically distributed) random vars $X_1, X_2, X_3, \ldots$

• $X_i$ has $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \text{Var}[X_i]$

• Consider the empirical mean: $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

• For any $\epsilon > 0$, as $n \to \infty$

$$\Pr(|\bar{X} - \mu| > \epsilon) \to 0.$$
• For any $\epsilon > 0$, as $n \to \infty$

$$\Pr(|\bar{X} - \mu| > \epsilon) \to 0.$$ 

• Proof:

$$E[\bar{X}] = E\left[\frac{X_1 + \cdots + X_n}{n}\right] = \mu$$

$$\text{Var}[\bar{X}] = \text{Var}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{\sigma^2}{n}$$

• By Chebyshev inequality,

$$\Pr(|\bar{X} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \to 0$$
• Consider **i.i.d.** (independent, identically distributed) random vars $X_1, X_2, X_3, \ldots$

• $X_i$ has $\mu = E[X_i]$

\[ \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \]

\[
\Pr \left( \lim_{n \to \infty} \left( \frac{X_1 + \cdots + X_n}{n} \right) = \mu \right) = 1
\]

• Strong Law $\Rightarrow$ Weak Law (but not vice versa)

• Strong law implies that for any $\epsilon > 0$, only finite number of $n$ such that weak law condition $|\overline{X} - \mu| \geq \epsilon$ is violated.
diffusion!
the law of large numbers

• Justifies the “frequency” interpretation of probability

• **NOT:**

• regression toward the mean

• **Gambler’s fallacy:** “I’m do for a win!”

• “Result will usually be close to the mean”

(http://stat-www.berkeley.edu/~stark/Java/Html/Iln.htm)
the central limit theorem (CLT)

• Consider \textit{i.i.d.} (independent, identically distributed) random vars $X_1, X_2, X_3, \ldots$

• $X_i$ has $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \text{Var}[X_i]$  

• As $n \to \infty$, 
\[
\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma \sqrt{n}} \to N(0, 1)
\]

• Restated: As $n \to \infty$, 
\[
\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)
\]
CLT in the real world
CLT in the real world

• CLT is the reason many things appear normally distributed
  • Many quantities are sums of (roughly) independent random vars

• Exam scores: sums of individual problems

• Election polling
  • Ask 100 people if they will vote for candidate X
  • Repeat this process with different groups to get empirical probabilities $p_1, p_2, \ldots, p_n$
  • Can produce a “confidence interval”
    • How likely is it that estimate for true $p$ is correct?
    • (We’ll do this next time)
CLT convergence
CLT convergence
in the real world…
in the real world…
in the real world...

Histogram of Daily Trading-Related Revenue* — Twelve Months Ended December 31, 2007

*Excludes daily profits and losses in the ABS CDO market, including recent subprime-related losses.
in the real world...
Chernoff bound was CLT (for binomial) in disguise

• Suppose $X \sim \text{Bin}(n,p)$

• $\mu = E[X] = pn$

• Chernoff bound:

For any $\delta$ with $0 < \delta < 1$,

\[ P(X > (1 + \delta)\mu) \leq e^{-\frac{\delta^2 \mu}{2}} \]

\[ P(X < (1 - \delta)\mu) \leq e^{-\frac{\delta^2 \mu}{3}} \]
• $X$ is a normal random variable $X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$E[X] = \mu \quad \text{Var}[X] = \sigma^2$$
• X is a normal random variable \( X \sim N(\mu, \sigma^2) \)

• \( Z \sim N(0, 1) \) “standard (or unit) normal”
  • Use \( \Phi(z) \) to denote CDF, i.e.

\[
\Phi(z) = \Pr(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx
\]
  • no closed form 😞
• Roll 10 6-sided dice

• \( X = \) total value of all 10 dice

• Win if: \( X \leq 25 \) or \( X \geq 45 \)

• Roll...

\[
E[X] = 10E[X_i] = 10(3.5) = 35 \quad \text{Var}(X) = 10 \text{Var}(X_i) = 10 \frac{35}{12} = \frac{350}{12}
\]

\[
1 - P(25.5 \leq X \leq 44.5) = 1 - P\left(\frac{25.5 - 35}{\sqrt{350/12}} \leq \frac{X - 35}{\sqrt{350/12}} \leq \frac{44.5 - 35}{\sqrt{350/12}}\right)
\]

\[
\approx 1 - (2\Phi(1.76) - 1) \approx 2(1 - 0.9608) = 0.0784
\]