The Central Limit Theorem

Random samples, iid random variables

- **Definition:** A **random sample** of size n from a given distribution is a set of n independent r.v.'s $X_1, X_2, \dots, X_n$, each having the given distribution, with expectation $E(X_i) = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2$. Such a set of random variables is also called **independent, identically distributed (iid)**.
- **Sample sum:** $S = \sum_{i=1}^n X_i$. $E(S) = n\mu$, $\mathrm{Var}(S) = n\sigma^2$,
- **Sample mean:** $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. $E(\overline{X}) = \mu$, $\mathrm{Var}(\overline{X}) = \sigma^2/n$

Normal approximation, Central Limit Theorem

The Central Limit Theorem (CLT) says that the mean and the sum of a random sample of a large enough size[1] from an (essentially) arbitrary distribution have **approximately** normal distribution: Given a random sample $X_1, \dots, X_n$ with $\mu = E(X_i)$ and $\sigma^2 = \mathrm{Var}(X_i)$, we have:

- The **sample sum** $S = \sum_{i=1}^n X_i$ is approximately normal $N(n\mu, n\sigma^2)$.
  Equivalently, the standardized version of S, $S^* = \dfrac{S - n\mu}{\sigma\sqrt{n}}$, is approximately standard normal.
- The **sample mean** $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is approximately normal $N(\mu, \sigma^2/n)$.
  Equivalently, the standardized version of $\overline{X}$, $\overline{X}^* = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$, is approximately standard normal.

Sums of independent normal r.v.'s

In the case the X_i's are already normal, the associated sums and random samples are **exactly** (and not just approximately) normal, with the appropriate parameters:

- **Sums and differences of two normal r.v.'s:** If X_1 and X_2 are $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively, then $X_1 \pm X_2$ is normal $N(\mu_1 \pm \mu_2, \sigma_1^2 + \sigma_2^2)$. (Note the plus sign in the formula for the variance of $X_1 + X_2$. Also, note that it is the variances that add up, not the standard deviations.)
- **General linear combinations of normal r.v.'s:** If each X_i is normal $N(\mu_i, \sigma_i^2)$, then $\sum_{i=1}^n c_i X_i$ is normal $N(\mu, \sigma^2)$, where $\mu = \sum_{i=1}^n c_i\mu_i$ and $\sigma^2 = \sum_{i=1}^n c_i^2\sigma_i^2$.

[1] That is, a large enough value of n. The larger n is, the better the CLT approximation becomes, but greater than 25 is usually more than enough in practice, and in some cases the CLT works already well for single digit values of n.