The Distribution of the Sample Mean

Suppose that $X_1, X_2, \ldots, X_n$ are a simple random sample from some distribution with expected value $\mu$ and standard deviation $\sigma$.

The most important parameter of most distributions is the expected value $\mu$, and it is often estimated by the sample mean $\bar{X}$.

So the sampling distribution of the sample mean $\bar{X}$ plays a central role in estimating $\mu$.

Some aspects of that sampling distribution are known exactly, and for some others we have useful approximations for large $n$. 
Mean and Standard Deviation

For any $n \geq 1$, the sampling distribution of $\bar{X}$ has the properties:

$$E(\bar{X}) = \mu_{\bar{X}} = \mu;$$
$$V(\bar{X}) = \sigma^2_{\bar{X}} = \frac{\sigma^2}{n};$$

and hence the standard deviation of $\bar{X}$ is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$
Normal Populations

If $X_1, X_2, \ldots, X_n$ are a sample from a *normal* distribution, then for any $n \geq 1$, $\bar{X}$ is also normally distributed.

We already know its expected value is $\mu$ and its standard deviation is $\sigma/\sqrt{n}$, so

$$\bar{X} \sim N \left( \mu, \frac{\sigma^2}{n} \right).$$

In particular, for any $z$,

$$P \left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z \right) = \Phi(z).$$
Other Populations

If \( X_1, X_2, \ldots, X_n \) are a simple random sample from any distribution with expected value \( \mu \) and standard deviation \( \sigma \), then for large \( n \), \( \bar{X} \) is approximately normally distributed.

Again, we know its expected value is \( \mu \) and its standard deviation is \( \sigma / \sqrt{n} \), so

\[
\bar{X} \approx N \left( \mu, \frac{\sigma^2}{n} \right).
\]

The Central Limit Theorem states that the approximation holds in the limit:

\[
P \left( \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq z \right) \to \Phi(z) \text{ as } n \to \infty.
\]
Binomial Distribution

We can use the Central Limit Theorem to approximate the binomial distribution.

Suppose that $X_1, X_2, \ldots, X_n$ are the success indicators in a binomial experiment. That is, each is a Bernoulli variable with $P(X_i = 1) = p$ for some $0 < p < 1$.

Then

$$E(X_i) = p$$

and

$$V(X_i) = p(1 - p).$$
The Central Limit Theorem implies that for large $n$

\[ P \left( \frac{\bar{X} - p}{\sqrt{p(1-p)/n}} \leq z \right) \approx \Phi(z). \]

So, if $X = X_1 + X_2 + \cdots + X_n = n\bar{X}$, then

\[ P \left( \frac{X - np}{\sqrt{np(1-p)}} \leq z \right) \approx \Phi(z). \]

The approximation is improved by replacing $X - np$ by $X - np + 1/2$ (a continuity correction).
Distribution of a Linear Combination

The sample mean $\bar{X}$ and the sample total $n\bar{X}$ are both examples of a linear combination of $X_1, X_2, \ldots, X_n$.

A general linear combination is of the form

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

for some constants $a_1, a_2, \ldots, a_n$.

For example, $\bar{X}$ is the special case $a_i = 1/n, i = 1, 2, \ldots, n$. 
Mean and Variance
Suppose that

\[ E(X_i) = \mu_i \text{ and } V(X_i) = \sigma^2_i, \quad i = 1, 2, \ldots, n. \]

Then

\[ E(Y) = a_1 E(X_1) + a_2 E(X_2) + \cdots + a_n E(X_n) \]
\[ = a_1 \mu_1 + a_2 \mu_2 + \cdots + a_n \mu_n \]

and, if \( X_1, X_2, \ldots, X_n \) are uncorrelated,

\[ V(Y) = a_1^2 V(X_1) + a_2^2 V(X_2) + \cdots + a_n^2 V(X_n) \]
\[ = a_1^2 \sigma^2_1 + a_2^2 \sigma^2_2 + \cdots + a_n^2 \sigma^2_n. \]
If $X_1, X_2, \ldots, X_n$ are correlated, the variance becomes

$$V(Y) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \text{Cov}(X_i, X_j)$$

$$= \sum_{i=1}^{n} a_i^2 V(X_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_i a_j \text{Cov}(X_i, X_j)$$

$$= \sum_{i=1}^{n} a_i^2 \sigma_i^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_i a_j \sigma_i \sigma_j \rho_{i,j}.$$ 

In the last expression, we use the definition

$$\rho_{i,j} = \text{Corr}(X_i, X_j) = \frac{\text{Cov}(X_i, X_j)}{\sigma_i \sigma_j}.$$
The proofs of these results are straightforward if tedious, and depend on nothing more than the fact that if $g(X_1, X_2, \ldots, X_n)$ and $h(X_1, X_2, \ldots, X_n)$ are any two functions of $X_1, X_2, \ldots, X_n$, then

$$E[g(X_1, X_2, \ldots, X_n) + h(X_1, X_2, \ldots, X_n)]$$
$$= E[g(X_1, X_2, \ldots, X_n)] + E[h(X_1, X_2, \ldots, X_n)].$$

The earlier statements about $\bar{X}$, that

$$E(\bar{X}) = \mu \text{ and } V(\bar{X}) = \frac{\sigma^2}{n},$$

are just the special case for uncorrelated $X_1, X_2, \ldots, X_n$ and $a_i = 1/n, \mu_i = \mu, \sigma_i = \sigma, i = 1, 2, \ldots, n.$
Difference Between Two Variables

We often need to compare two measurements.

- For example,

\[ X_1 = \text{blood pressure before taking a medication} \]
\[ X_2 = \text{blood pressure 1 hour after taking medication} \]
\[ Y = X_2 - X_1 = \text{change in blood pressure}. \]

This is the special case \( n = 2, a_1 = -1, a_2 = 1. \)
So

\[ E(Y) = \mu_2 - \mu_1 \]

and

\[ V(Y) = \sigma_1^2 + \sigma_2^2 - 2 \rho_{1,2} \sigma_1 \sigma_2 \]

and, if \( \rho_{1,2} = 0 \),

\[ V(Y) = \sigma_1^2 + \sigma_2^2. \]

Note that when \( X_1 \) and \( X_2 \) are uncorrelated, the variances add, because \( a_1^2 = a_2^2 = 1 \).
Normal Variables

If $X_1, X_2, \ldots, X_n$ are independent and *normally* distributed, then any linear combination

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

is also normally distributed.

This general result includes as a special case the fact that $\bar{X}$ is normally distributed when $X_1, X_2, \ldots, X_n$ are independent and normally distributed.

A more general Central Limit Theorem states that $Y$ is *approximately* normally distributed when $n$ is large, provided no $a_iX_i$ contributes substantially to the sum.