Suppose that several random variables $X_1, X_2, \ldots, X_n$ are associated with the outcome of some experiment.

Then any function of them, such as their mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is also a random variable:

- that is, its value is unknown before the experiment is carried out, but is determined by the outcome of the experiment.
Statistic

Any quantity that can be calculated from the observed values of $X_1, X_2, \ldots, X_n$ without knowing their distributions is called a statistic.

- Their mean $\bar{X}$ is a statistic.
- If all the $X$s have expected value $\mu$, that is
  \[
  E(X_i) = \mu, \; i = 1, 2, \ldots, n,
  \]
  then $(\bar{X} - \mu)$ is a function of $X_1, X_2, \ldots, X_n$, but it is not a statistic, because we need to know $\mu$, a parameter of their distributions, to calculate it.
Sampling Distribution of a Statistic

A statistic is a random variable, and therefore has a probability distribution.

The probability distribution of a statistic is called its sampling distribution.

- To some extent, this is just terminology: a statistic is a random variable, and consequently has a probability distribution.
- We use the term sampling distribution to emphasize that statistics vary from sample to sample.
Example 5.19

$X_1, X_2, \ldots, X_6$ are measured strengths of 6 randomly chosen items. Each has the Weibull distribution with $\alpha = 2, \beta = 5$.

The pdf:

curve(dweibull(x, 2, 5), to = 15)

Simulate $X_1, X_2, \ldots, X_6$ and calculate $\bar{X}$ and $S$, the standard deviation:

```r
x = rweibull(6, 2, 5); c(x, mean = mean(x), sd = sd(x))
```

Repeat: every sample is different, and so are the mean and sd.
Simple Random Sample

Often, as in Example 5.19, the random variables $X_1, X_2, \ldots, X_n$ satisfy:

- They are independent;
- They all have the same distribution.

In this case, we say that they are a *(simple) random sample* from that distribution.

Many statistical problems can be solved by finding the sampling distribution of some statistic in the context of a simple random sample.
Deriving a Sampling Distribution

The simplest case is the binomial experiment, with

\[ X_i = \begin{cases} 
1 & \text{if } i^\text{th} \text{ trial is a success} \\
0 & \text{otherwise} 
\end{cases} \]

Then \( X = \sum X_i \sim \text{Bin}(n, p) \), and \( \bar{X} = X/n \), so

\[
P(\bar{X} = x/n) = P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, \ldots, n.
\]

In some similar special cases, we can find the distribution of a statistic of interest exactly.
Simulation

In many cases, such as the Weibull samples of Example 5.19, the sampling distributions of the mean and the standard deviation cannot be written in any simple form.

We can always approximate these distributions by simulation (a simple form of the Monte Carlo method):

```r
nsim <- 1000
n <- 6
x <- matrix(rweibull(nsim * n, 2, 5), nsim, n)
hist(apply(x, 1, mean))
hist(apply(x, 1, sd))
```
We created $n_{\text{sim}} = 1000$ samples, each of size $n = 6$, and calculated the mean and sd of each.

For any $x > 0$, the fraction of sample means that are $\leq x$ is close to $P(\bar{X} \leq x) = F_{\bar{X}}(x)$, and gets closer as $n_{\text{sim}}$ is made larger.

More generally, we can get a good approximation to any probability involving the sample mean, the sample sd, or any other statistic, by making $n_{\text{sim}}$ large enough.