Statistics and Their Distributions

Definition: A statistic is any quantity whose value can be calculated from sample data. Prior to obtaining data, there is uncertainty as to what value of any particular statistic will result. Therefore, a statistic is a random variable and will be denoted by an uppercase letter; a lower case letter is used to represent the calculated or observed value of the statistic.

Remarks: Because of this uncertainty, before the data becomes available we view each observation as a random variable and denote the sample by $X_1, X_2, \ldots, X_n$. (Uppercase letters for random variables.) The observed values of the samples are denoted by $x_1, x_2, \ldots, x_n$. 
2°/ The sample mean, regarded as a statistic (before a sample has been selected or an experiment has been carried out), is denoted by \( \bar{x} \); the calculated value of this statistic is \( \bar{x} \).

3°/ Similarly, the sample standard deviation, thought of as a statistic, is denoted by \( s \), and its computed value is \( s \).

4°/ Any statistic, being a random variable, has a probability distribution. The probability distribution of a statistic is referred to as its sampling distribution to emphasize that it describes how the statistic varies in value across all samples that might be selected.
5°/ The probability distribution of any particular statistic depends not only on the population distribution (normal, uniform, etc) and the sample size n but also on the method of sampling.

Random Samples

To describe a sampling method often encountered (at least approximately) in practice.

Definition: The rv's \( \bar{X}_1, \bar{X}_2, \ldots, \bar{X}_n \) are said to form a random sample of size \( n \) if

1°/ The \( \bar{x}_i \)'s are independent rv's.

2°/ Every \( \bar{X} \): has the same probability distribution.
Remarks: 10/ Conditions 10/ and 20/ say that the X_i's are independent and identically distributed (i.i.d).

20/ If sampling is either with replacement or from an infinite (conceptual) population, conditions 10/ and 20/ are satisfied exactly.

30/ If sampling is without replacement, yet the sample size n is much smaller than the population size N (n/N ≤ 5% in practice), conditions 10/ and 20/ are approximately satisfied.

40/ The virtue of this sampling method is that the probability distribution of any statistic can be more easily obtained than for any other sampling methods.
Deriving the Sampling Distribution of a Statistic

Example 6.2: (There are relatively few different \( \bar{x} \) values in the population).

- Population distribution

<table>
<thead>
<tr>
<th>( x )</th>
<th>40</th>
<th>45</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f(x) )</td>
<td>0.2</td>
<td>0.3</td>
<td>0.5</td>
</tr>
</tbody>
</table>

with \( \mu = 46.5 \)

\[
\sigma^2 = 15.25
\]

- Let \( x_1 \) and \( x_2 \) constitute a sample from the population distribution.
  (i.e., \( x_1 \) and \( x_2 \) are independent, each with the probability distribution shown in Table 6.1.

- Table 6.2: Outcomes, probabilities, and \( \bar{x} \) and \( s^2 \),
<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$n_i$</th>
<th>$f(x_i, n_i) = p(x) f(x_i)$</th>
<th>$\bar{x}$</th>
<th>$s^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>40</td>
<td>0.04 = 0.2 * 0.2</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>45</td>
<td>45</td>
<td>0.06 = 0.2 * 0.3</td>
<td>42.5</td>
<td>12.5</td>
</tr>
</tbody>
</table>

To obtain the probability distribution of $\bar{x}$, we must consider each possible value $\bar{x}$ and compute its probability.

- $\bar{x} = 40$ occurs only 1 time in the table with probability 0.04:
  \[
p(\bar{x} = 40) = p(\bar{x} = 40) = 0.04
  \]

- $\bar{x} = 42.5$ occurs twice in the table with probability 0.06 and 0.06:
  \[
p(\bar{x} = 42.5) = p(\bar{x} = 42.5) = 0.06 + 0.06 = 0.12
  \]

- $\bar{x} = 45$ occurs three times with probabilities 0.1, 0.09 and 0.1,
  \[
p(\bar{x} = 45) = p(\bar{x} = 45) = 0.1 + 0.09 + 0.1 = 0.29
  \]

- etc.
The complete sampling distribution of $\bar{X}$ and $S^2$

<table>
<thead>
<tr>
<th></th>
<th>40</th>
<th>42.5</th>
<th>45</th>
<th>47.5</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi_{\bar{X}}(x)$</td>
<td>0.04</td>
<td>0.12</td>
<td>0.29</td>
<td>0.3</td>
<td>0.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>12.5</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{S^2}(y)$</td>
<td>0.38</td>
<td>0.42</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.6</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.06</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>0.09</td>
<td>0.06</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.15</td>
<td></td>
</tr>
</tbody>
</table>

Then:

\[ \mu_{\bar{X}} = E(\bar{X}) = \sum x \phi_{\bar{X}}(x) = 40(0.04) + 42.5(0.12) + 45(0.29) + 47.5(0.3) + 50(0.25) = 46.5 = \mu_{\bar{X}} \]

\[ \sigma_{\bar{X}}^2 = V(\bar{X}) = \sum x^2 \phi_{\bar{X}}(x) - \mu_{\bar{X}}^2 \]

\[ = (40)^2(0.04) + (42.5)^2(0.12) + (45)^2(0.29) + (47.5)^2(0.3) + (50)^2(0.25) - (46.5)^2 = 7.625 = \frac{15.25}{2} = \frac{\sigma_{\bar{X}}^2}{2} \]

\[ \mu_{S^2} = E(S^2) = \sum s^2 p_{S^2}(s) = (1)(0.38) + (12.5)(0.42) + (50)(0.25) = 15.25 = \sigma_{S^2}^2 \]
That is, the $\overline{X}$ sampling distribution is centered at the population mean $\mu$, and the $S^2$ sampling distribution is centered at the population variance $\sigma^2$. 