Sampling Distributions

You have seen probability distributions of various types. The normal distribution is an example of a \textit{continuous} distribution that is often used for quantitative measures such as weights, heights, etc. The binomial distribution is an example of a \textit{discrete} distribution because the possible outcomes are only a “discrete” set of values, 0, 1, \ldots, n. The value of the binomial random variable is the number of “successes” out of a random sample of n trials, in which the probability of success on a particular trail is $\pi$. 
The Binomial Distribution as a Sampling Distributions

The nature of the binomial distribution makes it a sampling distribution. In other words, the value of the binomial random variable is derived from a sample of size \( n \) from a population. You can think of the population as being a set containing 0’s and 1’s, with a 1 representing a “success” and 0 representing a “failure.” (The term “success” does not necessarily imply something good, nor does “failure” imply something bad.) The sample is obtained as a sequence of \( n \) trials, with each trial being a draw from the population of 0’s and 1’s. (Such trials are called “Bernoulli” trials.)

On the first trial, you draw a value from \( \{0, 1\} \), with \( P(1)=\pi \). Then you draw again in the same way, and do this repeatedly a total of \( n \) times. So your sample will be a set such as \( \{1, 1, 0, 1, 0, 1\} \) in the case for \( n=6 \). This set has 4 1’s and 2 0’s, so the value of the binomial random variable is \( y=4 \). The value of the binomial distribution is the sum of the outcomes from the trials; that is, the number of 1’s.
The Binomial Distribution as a Sampling Distribution

Recall that the binomial probability formula is

\[ P(y \text{ successes in } n \text{ trials}) = \frac{n!}{y!(n-y)!} \pi^y (1 - \pi)^{n-y} \]

where

\[ \pi = \text{probability of success on single trial} \]

and

\[ n = \text{number or trials} \]

Mean of the binomial distribution: \( n\pi \)

Variance of the binomial distribution: \( n\pi(1 - \pi) \)

This probability is derived from the sampling distribution of the number of successes that would result from a very large number of samples.
The Binomial distribution as a Sampling Distribution

Consider a population that consists 60% of 1’s and 40% of 0’s. If you draw a value at random, you get a 1 with probability .6 and a 0 with probability .4. Such a draw would constitute a Bernoulli trial with $P(1) = .6$.

Suppose you draw a sample of size $n$ and add up the value you obtain. This would give you a binomial random variable. Now suppose you do this again and again for a very large number of samples. The conceptual results make up a conceptual population. Here are the histograms that correspond to the population for values of $n$ equal to 1, 2, 3, 6, 10, and 20.

Notice how the shapes of the histograms change as $n$ increases. The distributions become more “mound-shaped” and symmetric.
The Binomial Sampling Distribution
and Statistical Inference

Here are the values of $P(y=k)$ and $P(y \leq k)$ for $n=20$ and $k=6-17$ ($P(y=k)<.005$ for $n<6$ or $n>17$):

<p>| | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td></td>
</tr>
</tbody>
</table>

$P(y=k)$

<p>| | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>.005</td>
<td>.015</td>
<td>.035</td>
<td>.071</td>
<td>.117</td>
<td>.160</td>
<td>.180</td>
<td>.165</td>
<td>.075</td>
<td>.035</td>
<td>.012</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$P(y \leq k)$

<p>| | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>.006</td>
<td>.021</td>
<td>.057</td>
<td>.128</td>
<td>.245</td>
<td>.404</td>
<td>.584</td>
<td>.750</td>
<td>.874</td>
<td>.949</td>
<td>.984</td>
<td>.996</td>
<td></td>
</tr>
</tbody>
</table>

Remember that these are the probabilities for $\pi=.6$.

Suppose you are sampling from a distribution with unknown $\pi$.

If you drew sample with $y=12$, would you have reason to doubt that $\pi=.6$? (Note: $y/n=.6$)

If you drew sample with $y=10$, would you have reason to doubt that $\pi=.6$? (Note: $y/n=.5$)

If you drew sample with $y=6$, would you have reason to doubt that $\pi=.6$? (Note: $y/n=.3$)
The Binomial Sampling Distribution
and Statistical Inference

Here are the numeric values of $P(y=k)$ for $n=10$, 3 decimal places:

\[
\begin{array}{cccccccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
\end{array}
\]

$P(y=k)$
\[
\begin{array}{cccccccccccc}
0.000 & 0.002 & 0.011 & 0.042 & 0.111 & 0.201 & 0.251 & 0.215 & 0.121 & 0.040 & 0.006 \\
\end{array}
\]

$P(y \leq k)$
\[
\begin{array}{cccccccccccc}
0.000 & 0.002 & 0.012 & 0.055 & 0.166 & 0.367 & 0.618 & 0.833 & 0.954 & 0.994 & 1.00 \\
\end{array}
\]

Remember that these are the probabilities for $\pi=.6$.

Suppose you are sampling from a distribution with unknown $\pi$.

If you drew sample with $y=6$, would you have reason to doubt that $\pi=.6$? (Note: $y/n=.6$)

If you drew sample with $y=5$, would you have reason to doubt that $\pi=.6$? (Note: $y/n=.5$)

If you drew sample with $y=3$, would you have reason to doubt that $\pi=.6$? (Note: $y/n=.3$)
Normal Distributions and Sampling Distributions

One of the most useful statistics is the sample mean, $\bar{y}$. Statistical inference based on $\bar{y}$ is derived from its sampling distribution. If the population from which the sample was obtained is normally distributed with mean $\mu$ and variance $\sigma^2$, then the sampling distribution of $\bar{y}$ is normal. Moreover, the mean of the sampling has mean $\mu$ and variance $\sigma^2/n$.

If $y_i \sim N(\mu, \sigma^2)$, $i=1,\ldots,n$, then $\bar{y} \sim N(\mu, \sigma^2 / n)$. 
Normal Distributions and Sampling Distributions

If the distribution from which the samples were obtained is not normal, then the sampling distribution of $\bar{Y}$ is only *approximately* normal, but the distribution becomes more nearly normal as $n$ increases. This is the *Central limit Theorem*.

**Central Limit Theorem:** If $Y$ is a random variable with mean $\mu$ and variance $\sigma^2$, then the sampling distribution of $\bar{Y}$ is approximately normal with mean $\mu$ and variance $\sigma^2/n$. 