NORMAL APPROXIMATION: THE CENTRAL LIMIT THEOREM

Everybody believes in the [normal approximation], the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact.”

G. Lippmann (French Physicist, 1845-1921)

- Change of scale for variance and expected values.
- Variance of binomial and proportions.
- Sample average as a random variable
  - Its mean: unbiased
  - Its variance: decreasing with sample size.
  - Its distribution: given by the CLT.
- Normal Approximation for Sample Means & Central Limit Theorem

Reminders about change of scales

If $X_1$ and $X_2$ indep. $var(X_1+X_2) = var(X_1)+var(X_2)$  
$Y = 2X$, $E(Y) = 2E(X)$, $var(Y) = 4var(X)$

$Y = aX$, $E(Y) = aE(X)$, $var(Y) = a^2var(X)$

Sums and averages of random variables

Suppose we have $n$ independent, identically distributed (we call these iid) Bernouilli’s $S = \sum_{i=1}^n X_i$ is a Binomial random variable with $E(S) = np$, $var(S) = np$, $var(S/n) = \frac{p}{n}$. Consider the the average of the $X_i$’s $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, how does it vary?

Example 1:
Score 1 if the voter votes for Kerry, score 0 if he/she votes for Bush. Suppose the overall population proportion of those in favor of Kerry is 0.5, and we want to estimate that unknown parameter by polls:

```r
round(runif(50))
```

```
1 0 1 0 0 1 0 1 1 0 1 1 1 1 1 1 1 0
1 0 0 0 1 0 0 1 1 1 1 1 0 1 0 0 1 0
```

\[
\hat{p} = \frac{\sum X}{n} = 30/50 = 0.6
\]

```r
> rbinom(1,50,.5)
[1] 26
> rbinom(1,50,.5)
[1] 29
> rbinom(1,50,.5)
[1] 25
> rbinom(1,50,.5)
[1] 23
```

\[
\hat{p}_2 = 0.52, \hat{p}_3 = 0.59, \hat{p}_4 = 0.59, \hat{p}_5 = 0.50, \hat{p}_6 = 0.46
\]
Sample Quantiles

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Frequency</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.950</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.900</td>
<td>0.001</td>
<td>0.000</td>
</tr>
<tr>
<td>0.850</td>
<td>0.004</td>
<td>0.000</td>
</tr>
<tr>
<td>0.800</td>
<td>0.008</td>
<td>0.000</td>
</tr>
<tr>
<td>0.750</td>
<td>0.011</td>
<td>0.000</td>
</tr>
<tr>
<td>0.700</td>
<td>0.014</td>
<td>0.000</td>
</tr>
<tr>
<td>0.650</td>
<td>0.017</td>
<td>0.000</td>
</tr>
<tr>
<td>0.600</td>
<td>0.020</td>
<td>0.000</td>
</tr>
<tr>
<td>0.550</td>
<td>0.023</td>
<td>0.000</td>
</tr>
<tr>
<td>0.500</td>
<td>0.026</td>
<td>0.000</td>
</tr>
<tr>
<td>0.450</td>
<td>0.029</td>
<td>0.000</td>
</tr>
<tr>
<td>0.400</td>
<td>0.032</td>
<td>0.000</td>
</tr>
<tr>
<td>0.350</td>
<td>0.035</td>
<td>0.000</td>
</tr>
<tr>
<td>0.300</td>
<td>0.038</td>
<td>0.000</td>
</tr>
<tr>
<td>0.250</td>
<td>0.041</td>
<td>0.000</td>
</tr>
<tr>
<td>0.200</td>
<td>0.044</td>
<td>0.000</td>
</tr>
<tr>
<td>0.150</td>
<td>0.047</td>
<td>0.000</td>
</tr>
<tr>
<td>0.100</td>
<td>0.050</td>
<td>0.000</td>
</tr>
<tr>
<td>0.050</td>
<td>0.053</td>
<td>0.000</td>
</tr>
<tr>
<td>0.000</td>
<td>0.056</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[
P\{0.25 \leq \hat{p} \leq 0.35\} = 0.179 + 0.192 + 0.164 = 0.53
\]

\[
\frac{n}{P\{0.25 \leq \hat{p} \leq 0.35\}}
\]

- \[
\begin{array}{r}
20 & .53 \\
40 & .61 \\
80 & .73 \\
400 & .97 \\
\end{array}
\]

\[
\begin{figure}
\text{Histogram of rbinom(20000, 20, 0.3)/20}
\end{figure}
\]

\[
\begin{figure}
\text{Histogram of rbinom(20000, 80, 0.3)/80}
\end{figure}
\]

\[
\begin{figure}
\text{Histogram of rbinom(20000, 40, 0.3)/40}
\end{figure}
\]
Sampling Distribution of $\bar{X}$

When we draw samples at random from a population $(\mu, \sigma^2)$ we obtain iid random variables forming a sample $X^\prime = \{X_\infty, X_\in, \ldots, X_n\}$ whose mean is $\bar{X} = \frac{1}{n} \sum^n X_i$. Sampling variability of $\bar{X}$ is described by the sampling distribution.

- Mean: The mean of the sampling distribution of $\bar{X}$ is the population mean $\mu$: $\mu_{\bar{X}} = \mu$.
- Standard deviation
  \[ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \]
- Shape: (a) If the population is Normal, then the sampling distribution of $\bar{X}$ is always Normal, regardless of the sample size $n$.
  (b) Central Limit Theorem: If $n$ is large then the sampling distribution of $\bar{Y}$ is approximately normal, even if the population is not Normal.

Normal Approximation to the Binomial

Sum of many independent 0/1 components with probabilities equal $p$ such that $npq \geq 3$.

Continuity Correction:

\[
P(a \leq X \leq b) \approx P\left( \frac{a - \frac{1}{2} - np}{\sqrt{np(1 - p)}} \leq Z \leq \frac{b + \frac{1}{2} - np}{\sqrt{np(1 - p)}} \right)
\]

Binomial Approximation to Normal Applet