Central Limit Theorem

Last time:
- Mean, Variance of a sum, average of i.i.d random variables.
- Also the PDF of a sum \((f_x \ast f_x \ast f_x \ast \ldots)\)

\[
\begin{align*}
\text{\textbullet} \ast \text{\textbullet} &= \text{\textbullet} \\
\text{\textbullet} \ast \text{\textbullet} &= \text{\textbullet}
\end{align*}
\]
The Central Limit Theorem (CLT) talks about what happens to the PDF/CDF of the sample mean as $n$ gets large.

Surprising result: no matter the distribution of $X$, in the limit, the distribution becomes Gaussian!
LAST TIME:

\[ X_i \text{ iid} \]

Mean \( \mu < \infty \)

Variance \( \sigma^2 < \infty \)

\[ S_n = X_1 + X_2 + \ldots + X_n \]

\[ M_n = \frac{1}{n} (X_1 + X_2 + \ldots + X_n) \]

\[ E(S_n) = n \mu \quad E(M_n) = \mu \]

\[ \text{Var}(S_n) = n \sigma^2 \quad \text{Var}(M_n) = \frac{1}{n} \sigma^2 \]
Consider a related r.v.

\[ Z_n = \frac{S_n - n\mu}{\sigma \sqrt{n}} \]

\[ E(Z_n) = 0 \]

\[ \text{Var}(Z_n) = \frac{\frac{1}{\sigma^2_n}}{n} \cdot \text{Var}(S_n) = 1 \]

No matter what \( n \) is, \( Z_n \) is zero mean unimodal variance.
The Central Limit Theorem says: CDF of $Z_n$

$$\lim_{n \to \infty} P(Z_n \leq u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-x^2/2} \, dx$$

This is the CDF of a Gaussian with mean 0, variance 1.

$$f_{\text{gaussian}}(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2\sigma^2} (x-\mu)^2}$$
$$\lim_{n \to \infty} F_{Z_n}(u) = \Phi(\frac{u}{\sqrt{n}})$$

CDF of $$Z_n$$

CDF of a "standard" Gaussian with $$\mu = 0, \sigma = 1$$.

$$= 1 - Q(u)$$

$$1 - \Phi(\frac{u}{\sqrt{n}})$$
No matter the CDF of $X_i$. 

Bernoulli 

Binomial 

Weird 

Psycho
For a continuous PDF,

PDF of \( Z_n \)

For a discrete/ mixed PMF
Implication: When \( n \) is big, we can use the CLT to get good approximations of probabilities using a table.
SPECIFICALLY:

\[ P(Z_n \leq u) = \phi(u) \]
\[ = 1 - Q(u) \]
\[ P\left( \frac{S_n - \eta u}{\sigma \sqrt{n}} \leq u \right) = \phi(u) \]
\[ = P(S_n - \eta u \leq u \sigma \sqrt{n}) \]
\[ = P(M_n - \mu \leq \frac{u \sigma}{\sqrt{n}}) \]
\[ P \left( \frac{S_n - \eta \mu}{\sigma \sqrt{n}} \leq z \right) = \phi(z) \]

\[ P \left( S_n - \eta \mu \leq z \sigma \sqrt{n} \right) = \phi(z) \]

\[ P \left( M_n - \mu \leq \frac{z \sigma}{\sqrt{n}} \right) = \phi(z) \]

\[ P \left( |M_n - \mu| \leq \frac{z \sigma}{\sqrt{n}} \right) \approx 1 - 2Q(z) \]
EX \[ X_1, X_2, \ldots, X_n \]

IID Bernoulli Variables

\[ P = \frac{1}{2}, \]
\[ M = \frac{1}{2}, \quad \sigma = \frac{1}{2}, \quad \sigma^2 = \frac{1}{4}. \]

How close is \( M_n \) to \( \frac{1}{2} \)?

Say \( n = 10,000 \)

\[ z = 1, \quad 0.005 \]

\[ P \left( \left| M_{10000} - \frac{1}{2} \right| \leq \frac{1}{200} \right) = \]

\[ 1 - 2Q(1) = 1 - 2(0.159) \]

\[ z = 0.632 \]
For what value of $n$ is

$$P \left( \left| M_n - \frac{1}{2} \right| \leq \frac{1}{200} \right) \geq 0.9$$

$$1 - 2Q(z) = 0.9$$

$$Q(z) = 0.05$$

$$z \approx 1.65$$

$$\frac{1}{200} = \frac{z \sigma}{\sqrt{n}} = \frac{(1.65)\left(\frac{1}{2}\right)}{\sqrt{n}}$$

$$\sqrt{n} = 200 \left(1.65\right)\left(\frac{1}{2}\right) \approx 165$$

$$n = (165)^2 = 27225.$$
A DIFFERENT SPIN:

SAY WE HAVE 10000 SAMPLES AND WE WANT TO KNOW THE CONFIDENCE INTERVAL OF $M_n$'S APPROXIMATION TO $M$.

FOR WHAT INTERVAL CAN WE BE 90% CONFIDENT?

$$P\left(\left|M_{10000} - \frac{1}{2}\right| \leq ??\right) \approx 0.9$$

$Z = 1.65$
\[
?? = \frac{Z_0}{\sqrt{n}} = \frac{(1.65) (\frac{1}{2})}{\sqrt{10000}}
\]

\[= 0.00825\]

\[P(1M_n - \frac{1}{2} \leq 0.00825) \approx 0.9\]

\[P( M_n \in \left[ \frac{1}{2} - 0.00825, \frac{1}{2} + 0.00825 \right] ) \approx 0.9.\]
When \( n \) is large, CLT will give much better approximations than Markov, Chebyshev, etc.

**EX**

Toss a fair die 20 times. Add up dots. What is \( P(\text{sum of dots is in } [60, 80]) \)?

For \( X \), uniform PMF on \([1, 2, \ldots, 6]\)

\[
\mu = \frac{1 + 6}{2} = 3.5
\]

\[
\sigma^2 = \frac{35}{12} = 2.92
\]

\( \sigma = 1.71 \).
Cheatyshev / Weak Law of Large Numbers:

\[ P(|S_{20} - 70| \leq 10) \]

\[ = P\left(|M_{20} - 3.5| \leq \frac{10}{20} = \frac{1}{2}\right) \]

\[ \geq 1 - \frac{\sigma^2 = 2.92}{20 \cdot \left(\frac{1}{2}\right)^2} = \frac{5}{12} \]

\[= 0.42 \]
ON THE OTHER HAND, 
THE CLT WOULD TELL US.

\[ P \left( \left| M_{20} - 3.5 \right| \leq \frac{1}{2} \right) \approx 1 - 2Q(z) \]

\[ \frac{Z \sigma}{\sqrt{n}} \]

\[ \frac{Z \sigma}{\sqrt{n}} = \frac{1}{2} \]

\[ Z(1.71) = \frac{1}{2} \]

\[ \frac{Z}{\sqrt{20}} = \frac{1}{2} \quad Z = 1.31 \]

\[ 1 - 2Q(1.31) = 1 - 2(0.095) = 0.81 \]
In real world estimation problems, we often want to estimate an unknown parameter by averaging a bunch of samples.

\[ X_i \text{ i.i.d.} \]

\[ M_n = \bar{X}_n = \frac{1}{n} \sum_{j=1}^{n} X_j \]

In the limit:

\[ \bar{X}_n \to \mu = E(X_i) \]
LET $1 - \alpha$ BE SOME HIGH PROBABILITY (e.g., 0.99)

$\alpha$ IS A SMALL NUMBER.

FIND $l$ AND $u$ S.T.

$$P(l(x) \leq \mu \leq u(x)) = 1 - \alpha$$

$$P(\mu \in [l(x), u(x)]) = 1 - \alpha$$

THAT IS, WE "KNOW" THAT $\mu$ IS IN A DETERMINED RANGE WITH HIGH PROBABILITY.

CONF. LEVEL $$(1 - \alpha)\%$ CONFIDENCE INTERVAL.
Suppose \( \mathbf{X}_j \) are iid

\[
\text{Unknown } \mu \\
\text{with known } \sigma
\]

\[
1 - 2Q(z) = P\left( -z \leq \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \leq z \right)
\]

\[
= P\left( -\frac{z \sigma}{\sqrt{n}} \leq \bar{X}_n - \mu \leq \frac{z \sigma}{\sqrt{n}} \right)
\]

\[
= P\left( \bar{X}_n - \frac{z \sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{z \sigma}{\sqrt{n}} \right)
\]

This interval contains \( \mu \) with probability \( 1 - 2Q(z) \)
Ex $X_j$ have unknown mean $\mu$ and variance $\sigma^2 = 1$.

Measure $X$ 100 times.

Find out that $\bar{X}_{100} = 5.25$.

Find the 95% confidence interval on $\mu$.

$n = 100$

$\sigma = 1$

$\alpha = 0.05 = 1 - 0.95$

$Z$ s.t. $\Theta(Z) = \frac{0.05}{2} = 0.025$

$Z_{0.025} = 1.96$ from table.
So if we set $\alpha$ and compute $z_{\alpha/2}$ so that

$$2\Phi(z) = \alpha \quad \Rightarrow \quad \Phi(z) = \frac{\alpha}{2}$$

$$1 - \alpha =$$

$$P\left( \mu \in X_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

95% The confidence interval for the mean is thus:

$$5.25 \pm 1.96 \cdot \frac{1}{\sqrt{100}}$$

$$5.25 \pm 0.196$$

$$= [5.05, 5.45]$$
IF WE DON'T KNOW $\sigma$
AND $n$ IS LARGE,
WE CAN STILL ESTIMATE
CONFIDENCE INTERVALS ON $\mu$
USING THE
STUDENT'S $t$ DISTRIBUTION.

OR IF WE KNOW $\mu$
AND WANT TO FIND CONFIDENCE
INTERVALS ON $\sigma^2$, WE
CAN USE LOOKUP TABLES
BASED ON THE $\chi^2$ (CHI-
SQUARED) DISTRIBUTION.
<table>
<thead>
<tr>
<th>$z$</th>
<th>$Q(z)$</th>
<th>$z$</th>
<th>$Q(z)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>5.000e-01</td>
<td>3.0</td>
<td>1.350e-03</td>
</tr>
<tr>
<td>0.1</td>
<td>4.602e-01</td>
<td>3.1</td>
<td>9.677e-04</td>
</tr>
<tr>
<td>0.2</td>
<td>4.207e-01</td>
<td>3.2</td>
<td>6.872e-04</td>
</tr>
<tr>
<td>0.3</td>
<td>3.821e-01</td>
<td>3.3</td>
<td>4.835e-04</td>
</tr>
<tr>
<td>0.4</td>
<td>3.446e-01</td>
<td>3.4</td>
<td>3.370e-04</td>
</tr>
<tr>
<td>0.5</td>
<td>3.085e-01</td>
<td>3.5</td>
<td>2.327e-04</td>
</tr>
<tr>
<td>0.6</td>
<td>2.743e-01</td>
<td>3.6</td>
<td>1.591e-04</td>
</tr>
<tr>
<td>0.7</td>
<td>2.420e-01</td>
<td>3.7</td>
<td>1.078e-04</td>
</tr>
<tr>
<td>0.8</td>
<td>2.119e-01</td>
<td>3.8</td>
<td>7.237e-05</td>
</tr>
<tr>
<td>0.9</td>
<td>1.841e-01</td>
<td>3.9</td>
<td>4.812e-05</td>
</tr>
<tr>
<td>1.0</td>
<td>1.587e-01</td>
<td>4.0</td>
<td>3.17e-05</td>
</tr>
<tr>
<td>1.1</td>
<td>1.357e-01</td>
<td>4.5</td>
<td>3.40e-06</td>
</tr>
<tr>
<td>1.2</td>
<td>1.151e-01</td>
<td>5.0</td>
<td>2.87e-07</td>
</tr>
<tr>
<td>1.3</td>
<td>9.680e-02</td>
<td>5.5</td>
<td>1.90e-08</td>
</tr>
<tr>
<td>1.4</td>
<td>8.076e-02</td>
<td>6.0</td>
<td>9.87e-10</td>
</tr>
<tr>
<td>1.5</td>
<td>6.681e-02</td>
<td>6.5</td>
<td>4.02e-11</td>
</tr>
<tr>
<td>1.6</td>
<td>5.480e-02</td>
<td>7.0</td>
<td>1.28e-12</td>
</tr>
<tr>
<td>1.7</td>
<td>4.457e-02</td>
<td>7.5</td>
<td>3.19e-14</td>
</tr>
<tr>
<td>1.8</td>
<td>3.593e-02</td>
<td>8.0</td>
<td>6.22e-16</td>
</tr>
<tr>
<td>1.9</td>
<td>2.872e-02</td>
<td>8.5</td>
<td>9.48e-18</td>
</tr>
<tr>
<td>2.0</td>
<td>2.275e-02</td>
<td>9.0</td>
<td>1.13e-19</td>
</tr>
<tr>
<td>2.1</td>
<td>1.786e-02</td>
<td>9.5</td>
<td>1.05e-21</td>
</tr>
<tr>
<td>2.2</td>
<td>1.390e-02</td>
<td>10.0</td>
<td>7.62e-24</td>
</tr>
<tr>
<td>2.3</td>
<td>1.072e-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.4</td>
<td>8.198e-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>6.210e-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.6</td>
<td>4.661e-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.7</td>
<td>3.467e-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.8</td>
<td>2.555e-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.9</td>
<td>1.866e-03</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>