Lecture 23: More on CLT + Parameter Estimation I
The CLT – Recap

**Theorem. (Central Limit Theorem)** The CDF of $Y_n$ converges to the CDF of the standard normal $N(0, 1)$, i.e.,

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

$$
Y_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma \sqrt{n}}
$$

$X_1, \ldots, X_n$ iid with mean $\mu$ and variance $\sigma^2$

One main application: (Normal) approximation of probabilities
Example – Recap

We flip $n$ independent coins, heads with probability $p = 0.75$.

$X = \# \text{ heads} \quad \mu = \mathbb{E}(X) = 0.75n \quad \sigma^2 = \text{Var}(X) = 0.1875n$

$$\mathbb{P}(X \leq 0.7n)$$

<table>
<thead>
<tr>
<th>$n$</th>
<th>exact</th>
<th>$\mathcal{N}(\mu, \sigma^2)$ approx</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.4744072</td>
<td>0.357500327</td>
</tr>
<tr>
<td>20</td>
<td>0.38282735</td>
<td>0.302788308</td>
</tr>
<tr>
<td>50</td>
<td>0.25191886</td>
<td>0.207108089</td>
</tr>
<tr>
<td>100</td>
<td>0.14954105</td>
<td>0.124106539</td>
</tr>
<tr>
<td>200</td>
<td>0.06247223</td>
<td>0.051235217</td>
</tr>
<tr>
<td>1000</td>
<td>0.00019359</td>
<td>0.000130365</td>
</tr>
</tbody>
</table>
Example – Bad Approximation

Fair coin flipped (independently) 40 times. Probability of 20 or 21 heads?

**Exact.** \( \mathbb{P}(X \in \{20, 21\}) = \left[ \binom{40}{20} + \binom{40}{21} \right] \left( \frac{1}{2} \right)^{40} \approx 0.2448 \)

**Approx.** \( \mathbb{P}(20 \leq X \leq 21) = \Phi \left( \frac{20 - 20}{\sqrt{10}} \leq \frac{X - 20}{\sqrt{10}} \leq \frac{21 - 20}{\sqrt{10}} \right) \)
\[
\approx \Phi \left( 0 \leq \frac{X - 20}{\sqrt{10}} \leq 0.32 \right)
\]
\[
= \Phi(0.32) - \Phi(0) \approx 0.1241
\]
Example – Even Worse Approximation

Fair coin flipped (independently) 40 times. Probability of 20 heads?

Exact. \[ P(X = 20) = \binom{40}{20} \left( \frac{1}{2} \right)^{40} \approx 0.1254 \]

Approx. \[ P(20 \leq X \leq 20) = 0 \]
Solution – Continuity Correction

Round to next integer!

To estimate probability that discrete RV lands in (integer) interval \(\{a, ..., b\}\), compute probability continuous approximation lands in interval \([a - \frac{1}{2}, b + \frac{1}{2}]\)
Example – Continuity Correction

Fair coin flipped (independently) 40 times. Probability of 20 or 21 heads?

Exact. \( \mathbb{P}(X \in \{20, 21\}) = \left( \binom{40}{20} + \binom{40}{21} \right) \left( \frac{1}{2} \right)^{40} \approx 0.2448 \)

Approx. \( \mathbb{P}(19.5 \leq X \leq 21.5) = \Phi \left( \frac{19.5 - 20}{\sqrt{10}} \leq \frac{X - 20}{\sqrt{10}} \leq \frac{21.5 - 20}{\sqrt{10}} \right) \approx \Phi \left( -0.16 \leq \frac{X - 20}{\sqrt{10}} \leq 0.47 \right) \)

\[ = \Phi(-0.16) - \Phi(0.47) \approx 0.2452 \]
Example – Continuity Correction

Fair coin flipped (independently) 40 times. Probability of 20 heads?

Exact. \( \mathbb{P}(X = 20) = \binom{40}{20} \left( \frac{1}{2} \right)^{40} \approx 0.1254 \)

Approx. \( \mathbb{P}(19.5 \leq X \leq 21.5) = \Phi \left( \frac{19.5 - 20}{\sqrt{10}} \leq \frac{X - 20}{\sqrt{10}} \leq \frac{20.5 - 20}{\sqrt{10}} \right) \)

\[ \approx \Phi \left( -0.16 \leq \frac{X - 20}{\sqrt{10}} \leq 0.16 \right) \]

\[ = \Phi(-0.16) - \Phi(0.16) \approx 0.1272 \]
**Theorem. (Central Limit Theorem)** The CDF of $Y_n$ converges to the CDF of the standard normal $\mathcal{N}(0,1)$, i.e.,

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

$$Y_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma \sqrt{n}}$$

$X_1, \ldots, X_n$ iid with mean $\mu$ and variance $\sigma^2$

**Theorem. (Weak Law of Large Numbers)** Let $X_1, \ldots, X_n$ iid with mean $\mu < \infty$ and variance $\sigma^2 < \infty$. Then,

$$\mathbb{P} \left( \left| \frac{X_1 + \cdots + X_n}{n} - \mu \right| \geq \epsilon \right) \to 0 \text{ as } n \to \infty$$

Proof: Use Chebyshev
Next: Learning from data
**Parameter Estimation – Workflow**

\[ \mathbb{P}(x|\theta) \rightarrow \text{Independent samples } x_1, \ldots, x_n \text{ from } \mathbb{P}(x|\theta) \rightarrow \text{Algorithm} \rightarrow \hat{\theta} \]

\( \theta = \text{unknown parameter} \)

**Example:** \( p(x|\theta) = \text{coin flip distribution with unknown } \theta = \text{probability of heads} \)

Observation: HTTHHHTHTTTTTHTHTTTTTHT

**Goal:** Estimate \( \theta \)
**Likelihood**

Say we see outcome $\text{HHTHH}$.  

$$\mathbb{P}(\text{HHTHH}|\theta) = \theta^4(1 - \theta)$$

Probability of observing the outcome $\text{HHTHH}$ if $\theta = \text{prob. of heads}$

$$\mathbb{P}(x|\theta) = \text{probability of (individual) outcome } x \text{ given model } \theta \text{ (H/T?)}$$

As a function of $x$ (fixed $\theta$): A probability  

$$\sum_x \mathbb{P}(x|\theta) = 1$$  

As a function of $\theta$ (fixed $x$): Likelihood
Likelihood of Different Observations

**Definition.** The **likelihood** of independent observations \( x_1, \ldots, x_n \) is

\[
L(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} \mathbb{P}(x_i | \theta)
\]

**Maximum Likelihood Estimation (MLE).** Given data \( x_1, \ldots, x_n \), find \( \hat{\theta} = \hat{\theta}(x_1, \ldots, x_n) \) ("the MLE") of model such that \( L(x_1, \ldots, x_n | \hat{\theta}) \) is maximized!

Usually: Solve \( \frac{\partial L(x_1, \ldots, x_n | \theta)}{\partial \theta} = 0 \) or \( \frac{\partial \ln L(x_1, \ldots, x_n | \theta)}{\partial \theta} = 0 \) [+check it’s a max!]

(Discrete case)
Example – Coin Flips

Coin-flip outcomes $x_1, \ldots, x_n$, with $n_H$ heads, $n_T$ tails

  - i.e., $n_H + n_T = n$

$L(x_1, \ldots, x_n | \theta) = \theta^{n_H} (1 - \theta)^{n_T}$

$\ln L(x_1, \ldots, x_n | \theta) = n_H \ln \theta + n_T \ln(1 - \theta)$

$\frac{\partial}{\partial \theta} \ln L(x_1, \ldots, x_n | \theta) = n_H \cdot \frac{1}{\theta} - n_T \cdot \frac{1}{1 - \theta}$

Goal: estimate $\theta = \text{prob. heads.}$

Solve $n_H \cdot \frac{1}{\theta} - n_T \cdot \frac{1}{1 - \theta} = 0$
The Continuous Case

Given \( n \) samples \( x_1, \ldots, x_n \) from a Gaussian \( \mathcal{N}(\mu, \sigma^2) \), estimate \( \theta = (\mu, \sigma^2) \)

**Definition.** The **likelihood** of independent observations \( x_1, \ldots, x_n \) is

\[
L(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} f(x_i | \theta)
\]

Density function! (Why?)
Why density?

• Density $\neq$ probability, but:
  – For maximizing likelihood, we really only care about relative likelihoods, and density captures that
  – has desired property that likelihood increases with better fit to the model
samples $x_1, ..., x_n \in \mathbb{R}$ from Gaussian $\mathcal{N}(\mu, 1)$. Most likely $\mu$?
[i.e., we are given the promise that the variance is one]
$n$ samples $x_1, \ldots, x_n \in \mathbb{R}$ from Gaussian $\mathcal{N}(\mu, 1)$. Most likely $\mu$?

$\mu = 0$?

Unlikely …
$n$ samples $x_1, ..., x_n \in \mathbb{R}$ from Gaussian $\mathcal{N}(\mu, 1)$. Most likely $\mu$?

$\mu = 3$?

Better, but optimal?
Example – Gaussian Parameters

Normal outcomes $x_1, \ldots, x_n$, known variance $\sigma^2 = 1$

Goal: estimate $\mu = \text{expectation}$

$$L(x_1, \ldots, x_n | \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2}} = \left( \frac{1}{\sqrt{2\pi}} \right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2}}$$

$$\ln L(x_1, \ldots, x_n | \mu) = -n \frac{\ln 2\pi}{2} - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2}$$
Example – Gaussian Parameters

Goal: estimate $\mu = \text{expectation}$

Normal outcomes $x_1, \ldots, x_n$, known variance $\sigma^2 = 1$

$$\ln L(x_1, \ldots, x_n | \mu) = -n \frac{\ln 2\pi}{2} - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2}$$

Note: $$\frac{\partial}{\partial \mu} \frac{(x_i - \mu)^2}{2} = \frac{1}{2} \cdot 2 \cdot (x_i - \mu) \cdot (-1) = \mu - x_i$$

$$\frac{\partial}{\partial \mu} \ln L(x_1, \ldots, x_n | \mu) = \sum_{i=1}^{n} (x_i - \mu) = \sum_{i=1}^{n} x_i - n\mu = 0$$

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

In other words, MLE is the sample mean of the data.
Next: \( n \) samples \( x_1, \ldots, x_n \in \mathbb{R} \) from Gaussian \( \mathcal{N}(\mu, \sigma^2) \). Most likely \( \mu \) and \( \sigma^2 \)?