today: - continue Bayesian approach
- MLE

(continuation from last class):

posterior belief $p(\theta | X=x)$ contains all the info from data that we need to answer new questions

E.g. question: what is probability of head (F=1) on next flip?

as frequentist: $p(F=1|\theta) = \hat{\theta}$

as a Bayesian: $p(F=1|X=x) = \int_{\theta} p(F=1, \theta | X=x) d\theta$

\[p(x_i, y(z)) = p(z | x_i, z) p(y | z)\]

\[p(\hat{\theta}, y | z, w) = p(z | y, z, w) p(y | z, w)\]

\[
\mathbb{E}[\theta | X=x] = \int_{\theta} p(\theta | X=x) d\theta = \mathbb{E}[\hat{\theta} | X=x]
\]

a meaningful Bayesian estimator of $\theta$
A meaningful Bayesian estimator of $\theta$

$$\hat{\theta}_{\text{Bayes}}(x) \triangleq \mathbb{E}[\theta | X=x] \quad \text{(posterior mean)}$$

(notation: $\hat{\theta}$ is observation $\rightarrow \theta$)

Our coin example:

$$p(\theta | X=x) = \text{Beta}(\theta | \alpha = x+1, \beta = n-x+1)$$

Prior

$$p(\theta) = \text{uniform}(\theta)$$

Thus

$$\mathbb{E}[\hat{\theta}_{\text{Bayes}}(x)] = \mathbb{E}[\theta | X=x] = \frac{x+1}{n+2}$$

This is biased

**In frequentist statistics**: Consider multiple possible estimators

- **MLE**
  - moment matching
  - Bayesian posterior mean

But asymptotically unbiased

$i.e.$

$$\mathbb{E}_X[\hat{\theta}_{\text{Bayes}}(x)] = \theta$$
and then analyze their statistical properties.

- biased?
- variance?
- consistent?

**Maximum Likelihood Principle**

setup: given a parametric family \( p(x; \theta) \) for \( \theta \in \Theta \)

- we want to estimate \( \theta \)

\[
\hat{\theta}_{ML}(x) = \arg \max_{\theta \in \Theta} p(x; \theta)
\]

\[
\Delta \hat{\theta}_{ML}(x) \subseteq \text{L}(\theta)
\]

\[
\text{"likelihood function" of } \theta
\]

*Example:* \( n \) coin flips \( x \)

\( x \sim \text{Bin}(n, \theta) \)

\[ p(x; \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} \]

**trick:** to maximize \( \log L(\theta) \) instead of \( L(\theta) \)
Justification: \( \log(\cdot) \) is strictly increasing

\[ \text{i.e. } a < b \iff \log a < \log b \]

\[ \Rightarrow \arg\max_{\theta \in \Theta} \log p(x; \theta) = \arg\max_{\theta \in \Theta} p(x; \theta) \]

\[ \log p(x; \theta) = \log \left( \frac{p(x)}{p(\theta)} \right) + x \log \theta + (n-x) \log (1-\theta) = 0(\theta) \]

\[ f'(\theta) = 0 \]

\[ f''(\theta) < 0 \]

\[ \text{look for } \theta \text{ s.t. } \left. \frac{\partial^2 f}{\partial \theta^2} \right|_{\theta=0} = 0 \]

\[ \text{i.e. } \frac{x}{\theta} - \frac{n-x}{1-\theta} = 0 \]

\[ x(1-\theta) - \theta(n-x) = 0 \]

\[ \Rightarrow \theta = \frac{x}{n} \]

\[ \text{here } \frac{\partial^2 f}{\partial \theta^2} \bigg|_{\theta=0} = \frac{x}{n} \]

\[ \text{i.e. relative frequency} \]

Some optimization comments:

- \( f'(\theta) = 0 \) is necessary condition for a local max when \( \theta \) is in interior of \( \Theta \)

- Also need to check \( f''(\theta) < 0 \) for a local max
also need to check $f''(c) < 0$ for a local min

$\nabla f(c) = 0$  

only local result in general

but if $f''(c) < 0 \forall c \in G$ , function is said "concave"

in this case, $f(c) = 0$ is sufficient for global max

be careful with boundary cases i.e. $c \in \text{boundary}(G)$

e.g.

another example $G = [0, 1]$  

Some notes about MLE

- does not always exist $[G \in \text{bd}(G)$ but $G$ is open] or when $G = +\infty$
  
e.g. $G = [0, 1]$

- is not necessary unique [i.e. multiple maxima]
Example 2: Multinomial distribution

Suppose $X_i$ is discrete R.V. on $K$ choices 

\[ \pi \in \Delta_K \]

We could choose \( \mathbb{E}X_i = \varepsilon_1, \ldots, K \varepsilon_K \) 

but instead, convenient to encode with unit basis in \( \mathbb{R}^K \)

i.e. \( \mathbb{E}X_i = \varepsilon_1, \ldots, K \varepsilon_K \) where \( \varepsilon_j \in \mathbb{R}^K \) "one hot encoding"

\[ \varepsilon_j = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \]

Parameter for discrete R.V. $X_i$: 

\[ \pi \in \Delta_K \]

\[ \Delta_K = \left\{ \pi \in \mathbb{R}^K : \pi_j > 0 \forall j, \frac{1}{K} \sum_{j=1}^{K} \pi_j = 1 \right\} \]

Probability simplex on $K$ choices

We will write $X_i \sim \text{Mult}(\pi)$ "Multinomial"

Consider $X_i \overset{iid}{\sim} \text{Mult}(\pi)$

Then \( \mathbb{E}X = \frac{1}{n} \sum_{i=1}^{n} X_i \sim \text{Mult}(n, \pi) \)

"Multinomial distribution"
\[ \Omega_{\chi} = \{ (\eta_1, \ldots, \eta_k) \mid \exists \eta_i \in \mathbb{N}, \sum_{i=1}^{k} n_i = n \} \]