Likelihood of Data

- Consider n I.I.D. random variables $X_1, X_2, \ldots, X_n$
  - X_i a sample from density function $f(X_i \mid \theta)$
    - Note: now explicitly specify parameter θ of distribution
  - We want to determine how "likely" the observed data $(x_1, x_2, \ldots, x_n)$ is based on density $f(X_i \mid \theta)$
  - Define the **Likelihood function**, $L(\theta)$:

$$L(\theta) = \prod_{i=1}^{n} f(X_i \mid \theta)$$

    - This is just a product since X_i are I.I.D.
  - Intuitively: what is probability of observed data using density function $f(X_i \mid \theta)$, for some choice of θ

Demo

Maximum Likelihood Estimator

- The **Maximum Likelihood Estimator** (MLE) of θ, is the value of θ that maximizes $L(\theta)$
  - More formally: $\theta_{MLE} = \arg\max_{\theta} L(\theta)$
  - More convenient to use **log-likelihood function**, $LL(\theta)$:

$$LL(\theta) = \log L(\theta) = \log \prod_{i=1}^{n} f(X_i \mid \theta) = \sum_{i=1}^{n} \log f(X_i \mid \theta)$$

  - Note that *log* function is "monotone" for positive values
    - Formally: $x \le y \Leftrightarrow \log(x) \le \log(y)$ for all $x, y > 0$
  - So, θ that maximizes $LL(\theta)$ also maximizes $L(\theta)$
    - Formally: $\arg\max_{\theta} LL(\theta) = \arg\max_{\theta} L(\theta)$
    - Similarly, for any positive constant c (not dependent on θ): $\arg\max_{\theta} (c \cdot LL(\theta)) = \arg\max_{\theta} LL(\theta) = \arg\max_{\theta} L(\theta)$

Computing the MLE

- General approach for finding MLE of θ
  - Determine formula for $LL(\theta)$
  - Differentiate $LL(\theta)$ w.r.t. (each) θ: $\frac{\partial LL(\theta)}{\partial \theta}$
  - To maximize, set $\frac{\partial LL(\theta)}{\partial \theta} = 0$
  - Solve resulting (simultaneous) equation to get θ_{MLE}
    - Make sure that derived $\hat{\theta}_{MLE}$ is actually a maximum (and not a minimum or saddle point). E.g., check $LL(\theta_{MLE} \pm \varepsilon) < LL(\theta_{MLE})$
      - This step often ignored in expository derivations
      - So, we'll ignore it here too (and won't require it in this class)
  - For many standard distributions, someone has already done this work for you. (Yay!)

Maximizing Likelihood with Bernoulli

- Consider I.I.D. random variables $X_1, X_2, \ldots, X_n$
  - $X_i \sim \text{Ber}(p)$
  - Probability mass function, $f(X_i \mid p)$, can be written as:

$$f(X_i \mid p) = p^{x_i}(1-p)^{1-x_i} \quad \text{where } x_i = 0 \text{ or } 1$$

  - Likelihood: $L(\theta) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$
  - Log-likelihood:

$$LL(\theta) = \sum_{i=1}^{n} \log(p^{X_i}(1-p)^{1-X_i}) = \sum_{i=1}^{n} [X_i(\log p) + (1-X_i)\log(1-p)]$$

$$= Y(\log p) + (n-Y)\log(1-p) \quad \text{where } Y = \sum_{i=1}^{n} X_i$$

  - Differentiate w.r.t. p, and set to 0:

$$\frac{\partial LL(p)}{\partial p} = Y\frac{1}{p} + (n-Y)\frac{-1}{1-p} = 0 \quad \Rightarrow \quad p_{MLE} = \frac{Y}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Maximizing Likelihood with Poisson

- Consider I.I.D. random variables $X_1, X_2, \ldots, X_n$
  - $X_i \sim \text{Poi}(\lambda)$
  - PMF: $f(X_i \mid \lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$ Likelihood: $L(\theta) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{X_i}}{X_i!}$
  - Log-likelihood:

$$LL(\theta) = \sum_{i=1}^{n} \log\left(\frac{e^{-\lambda}\lambda^{X_i}}{X_i!}\right) = \sum_{i=1}^{n} [-\lambda \log(e) + X_i \log(\lambda) - \log(X_i!)]$$

$$= -n\lambda + \log(\lambda)\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log(X_i!)$$

  - Differentiate w.r.t. λ, and set to 0:

$$\frac{\partial LL(\lambda)}{\partial \lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} X_i = 0 \quad \Rightarrow \quad \lambda_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Maximizing Likelihood with Normal

- Consider I.I.D. random variables $X_1, X_2, \ldots, X_n$
  - $X_i \sim N(\mu, \sigma^2)$
  - PDF: $f(X_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\mu)^2/(2\sigma^2)}$
  - Log-likelihood:

$$LL(\theta) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\mu)^2/(2\sigma^2)}\right) = \sum_{i=1}^{n} \left[-\log(\sqrt{2\pi}\sigma) - (X_i-\mu)^2/(2\sigma^2)\right]$$

  - First, differentiate w.r.t. μ, and set to 0:

$$\frac{\partial LL(\mu, \sigma^2)}{\partial \mu} = \sum_{i=1}^{n} 2(X_i-\mu)/(2\sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) = 0$$

  - Then, differentiate w.r.t. σ, and set to 0:

$$\frac{\partial LL(\mu, \sigma^2)}{\partial \sigma} = \sum_{i=1}^{n} -\frac{1}{\sigma} + 2(X_i-\mu)^2/(2\sigma^3) = -\frac{n}{\sigma} + \sum_{i=1}^{n}(X_i-\mu)^2/(\sigma^3) = 0$$

Being Normal, Simultaneously

- Now have two equations, two unknowns:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)=0 \qquad -\frac{n}{\sigma}+\sum_{i=1}^{n}(X_i-\mu)^2/(\sigma^3)=0$$

- First, solve for μ_{MLE}:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)=0 \;\Rightarrow\; \sum_{i=1}^{n}X_i=n\mu \;\Rightarrow\; \mu_{MLE}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

- Then, solve for σ^2_{MLE}:

$$-\frac{n}{\sigma}+\sum_{i=1}^{n}(X_i-\mu)^2/(\sigma^3)=0 \;\Rightarrow\; n\sigma^2=\sum_{i=1}^{n}(X_i-\mu)^2$$

$$\sigma^2_{MLE}=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_{MLE})^2$$

- Note: μ_{MLE} unbiased, but σ^2_{MLE} biased (same as MOM)

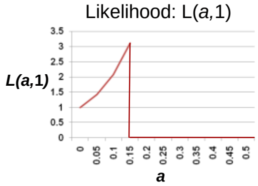
Maximizing Likelihood with Uniform

- Consider I.I.D. random variables $X_1, X_2, ..., X_n$
  - X_i ~ Uni(a, b)
  - PDF: $f(X_i \mid a,b)=\begin{cases}\frac{1}{b-a} & a < x_i < b \\ 0 & \text{otherwise}\end{cases}$
  - Likelihood: $L(\theta)=\begin{cases}\left(\frac{1}{b-a}\right)^n & a < x_1, x_2, ..., x_n < b \\ 0 & \text{otherwise}\end{cases}$
    - Constraint $a < x_1, x_2, \ldots, x_n < b$ makes differentiation tricky
    - Intuition: want interval size (b – a) to be as small as possible to maximize likelihood function for each data point
    - But need to make sure all observed data contained in interval
      - If all observed data not in interval, then $L(\theta) = 0$
  - Solution: $a_{MLE} = \min(x_1, \ldots, x_n)$ $b_{MLE} = \max(x_1, \ldots, x_n)$

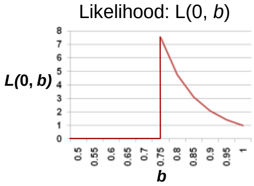
Understanding MLE with Uniform

- Consider I.I.D. random variables $X_1, X_2, ..., X_n$
  - X_i ~ Uni(0, 1)
  - Observe data:
    - 0.15, 0.20, 0.30, 0.40, 0.65, 0.70, 0.75

Likelihood: L(a,1)



Likelihood: L(0, b)



Once Again, Small Samples = Problems

- How do small samples effect MLE?
  - In many cases, $\mu_{MLE}=\frac{1}{n}\sum_{i=1}^{n}X_i$ = sample mean
    - Unbiased. Not too shabby…
  - As seen with Normal, $\sigma^2_{MLE}=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_{MLE})^2$
    - Biased. Underestimates for small n (e.g., 0 for n = 1)
  - As seen with Uniform, $a_{MLE} \geq a$ and $b_{MLE} \leq b$
    - Biased. Problematic for small n (e.g., a = b when n = 1)
  - Small sample phenomena intuitively make sense:
    - Maximum likelihood ⇒ best explain data we've seen
    - Does not attempt to generalize to unseen data

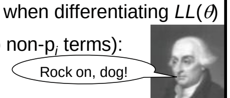
Properties of MLE

- Maximum Likelihood Estimators are generally:
  - Consistent: $\lim_{n\to\infty} P(|\hat{\theta}-\theta| < \varepsilon)=1$ for $\varepsilon > 0$
  - Potentially biased (though asymptotically less so)
  - Asymptotically optimal
    - Has smallest variance of "good" estimators for large samples
  - Often used in practice where sample size is large relative to parameter space
    - But be careful, there are some very large parameter spaces
    - Joint distributions of several variables can cause problems
      - Parameter space grows exponentially
      - Parameter space for 10 dependent binary variables ≈ 2^{10}

Maximizing Likelihood with Multinomial

- Consider I.I.D. random variables $Y_1, Y_2, ..., Y_n$
  - Y_k ~ Multinomial($p_1, p_2, ..., p_m$), where $\sum_{i=1}^{m} p_i = 1$
  - X_i = number of trials with outcome i where $\sum_{i=1}^{m} X_i = n$
  - PDF: $f(X_1,..., X_m \mid p_1,..., p_m)=\frac{n!}{x_1!x_2!\cdots x_m!}p_1^{x_1}p_2^{x_2}\ldots p_m^{x_m}$
  - Log-likelihood: $LL(\theta)=\log(n!)-\sum_{i=1}^{m}\log(X_i!)+\sum_{i=1}^{m}X_i\log(p_i)$
  - Account for constraint $\sum_{i=1}^{m} p_i = 1$ when differentiating $LL(\theta)$
  - Use Lagrange multipliers (drop non-p_i terms):

$$A(\theta)=\sum_{i=1}^{m}X_i\log(p_i)+\lambda\left(\sum_{i=1}^{m}p_i-1\right)$$



Joseph-Louis Lagrange (1736-1813)

Home on Lagrange

- Want to maximize:

$$A(\theta) = \sum_{i=1}^{m} X_i \log(p_i) + \lambda(\sum_{i=1}^{m} p_i - 1)$$

  - Differentiate w.r.t. each p_i, in turn:

$$\frac{\partial A(\theta)}{\partial p_i} = X_i \frac{1}{p_i} + \lambda = 0 \quad \Rightarrow \quad p_i = \frac{-X_i}{\lambda}$$

  - Solve for λ, noting $\sum_{i=1}^{m} X_i = n$ and $\sum_{i=1}^{m} p_i = 1$:

$$\sum_{i=1}^{m} p_i = \sum_{i=1}^{m} \frac{-X_i}{\lambda} \quad \Rightarrow \quad 1 = \frac{-n}{\lambda} \quad \Rightarrow \quad \lambda = -n$$

  - Substitute λ into p_i, yielding: $p_i = \frac{X_i}{n}$
  - Intuitive result: probability p_i = proportion of outcome *i*

When MLE's Attack!

- Consider 6-sided die
  - X ~ Multinomial(p_1, p_2, p_3, p_4, p_5, p_6)
  - Roll n = 12 times
  - Result: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes
  - Consider MLE for p_i:
    - p_1 = 3/12, p_2 = 2/12, p_3 = 0/12, p_4 = 3/12, p_5 = 1/12, p_6 = 3/12
  - Based on estimate, infer that you will ***never*** roll a three
  - Do you really believe that?
    - Frequentist: Need to roll more! Probability = frequency in limit
    - Bayesian: Have prior beliefs of probability, even before any rolls!

Need a Volunteer

Two Envelopes

- I have two envelopes, will allow you to have one
  - One contains \$X, the other contains \$2X
  - Select an envelope
    - Open it!
  - Now, would you like to switch for other envelope?
  - To help you decide, compute E[\$ in other envelope]
    - Let Y = \$ in envelope you selected

$$E[\$\text{ in other envelope}] = \frac{1}{2} \cdot \frac{Y}{2} + \frac{1}{2} \cdot 2Y = \frac{5}{4} Y$$

  - Before opening envelope, think either equally good
  - So, what happened by opening envelope?
    - And does it really make sense to switch?