Stat 111 Presentation Likelihood Function and Sufficient Statistics

Mathematical Statistics 2, Spring 2025 Caroline Yao

We first try to define the likelihood function $L(\theta)$ as the probability or probability density of the observed data values for a given value of the unknown parameter θ.
Suppose that random variables $X_1, \dots, X_n$ have a joint density or frequency function $f(x_1, x_2, \dots, x_n \mid \theta)$. Given observed values $X_i = x_i$, where $i = 1, \dots n$, the likelihood of θ as a function of $x_1, \dots x_n$ is defined as

$$L(\theta) = f(x_1, x_2, \dots, x_n \mid \theta)$$

It is important to notice that this is a function of θ rather than x_i's.

The **maximum likelihood estimate (mle)** of θ is that value of θ that maximizes the likelihood, which is what makes the observed data "most probable" or "most likely". So this is a natural estimate of θ. We often write it as $L(\hat{\theta})$.

We now move on to see the fact that very different likelihood functions could all give the same maximum likelihood estimates. Consider the case where we have $X_1, \dots X_n \ iid \sim N(\theta, 1)$, then we will have MLE is $\hat{\theta} = \bar{X}$. Now consider another set of random variables $X_1, \dots X_n \sim Unif(0, 2\theta)$. We know that the MLE is $\hat{\theta} = X_{(n)}/2$. But it is possible that $\bar{X} = X_{(n)}/2$.

We now consider the example where we have a discrete parameter space. Suppose we have a coin that is either fair or 2-headed. We want to find the likelihood function $L(\theta)$ following two independent trials with the coin landing heads 0, 1, or 2 times.

$$X_1 = \begin{cases} 1 & \text{trial 1 land Heads} \\ 0 & \text{otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{trial 2 land Heads} \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} \text{0 Heads case:} \quad & L(\theta) = \mathbb{P}(X_1 = 0, X_2 = 0; \theta) = (1-\theta)^2 \\ \text{1 Head case:} \quad & L(\theta) = \mathbb{P}(X_1 = 1, X_2 = 0; \theta) = \theta(1-\theta) \\ \text{2 Heads case:} \quad & L(\theta) = \mathbb{P}(X_1 = 1, X_2 = 1; \theta) = \theta^2 \end{aligned}$$

We see that for outcome with 2 heads,

$$L(0.5) = 0.5^2 = 0.25$$
$$L(1) = 1^2 = 1$$
$$\frac{L(1)}{L(0.5)} = \frac{1}{0.25} = 4$$

So it is 4 times more likely that $\theta = 1$ than $\theta = 0.5$.

We now consider $\theta \in [0, 1]$, a continuous probability parameter. We want to graph $L(\theta)$ and $l(\theta) = \log(L(\theta))$ for 0, 1, or 2 success in 2 independent trials.

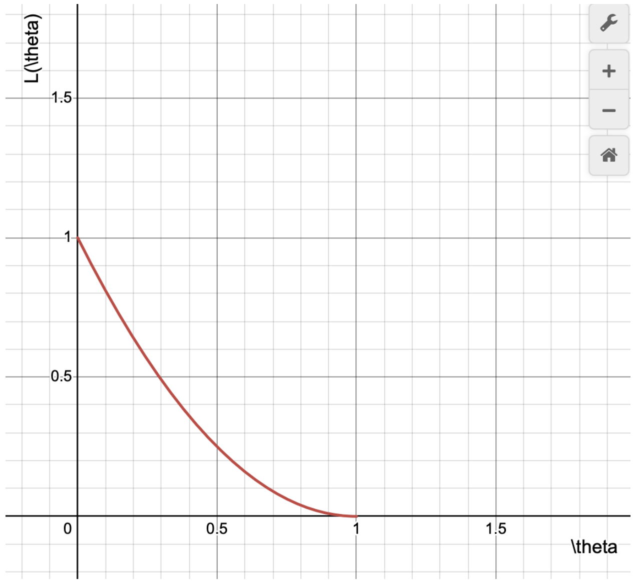

Figure 1: $L(\theta) = (1-\theta)^2$, $0 \le \theta \le 1$

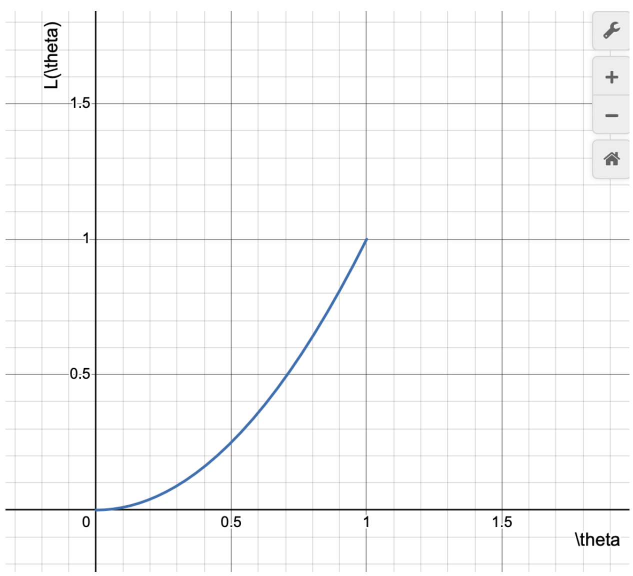

Figure 2: $L(\theta) = \theta^2$, $0 \le \theta \le 1$

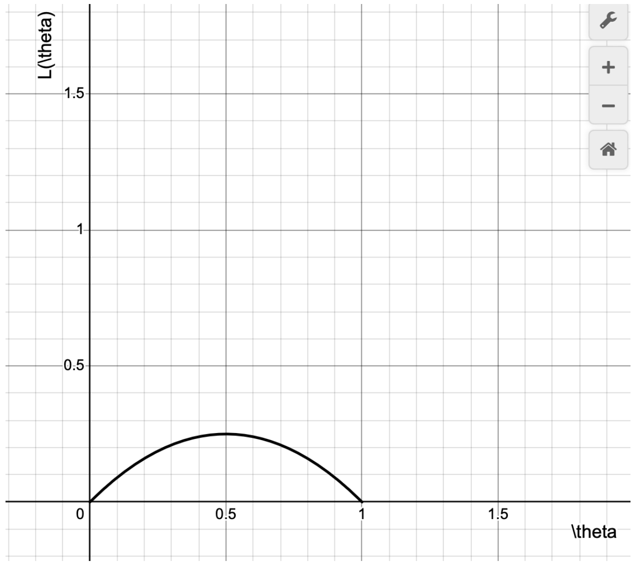

Figure 3: $L(\theta) = \theta(1-\theta)$, $0 \leq \theta \leq 1$

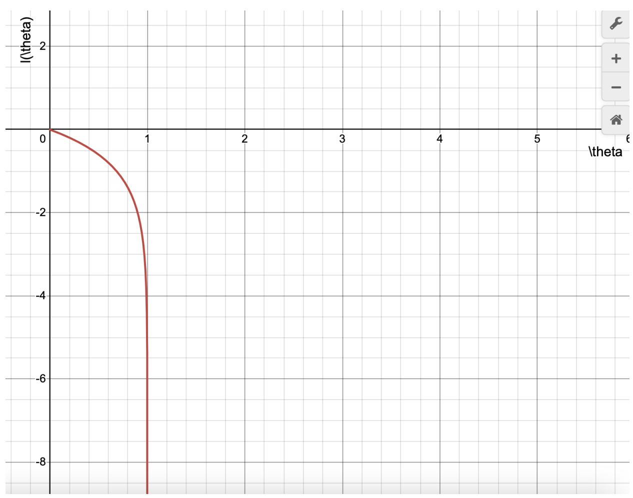

Figure 4: $l(\theta) = \log((1-\theta)^2)$, $0 \leq \theta \leq 1$

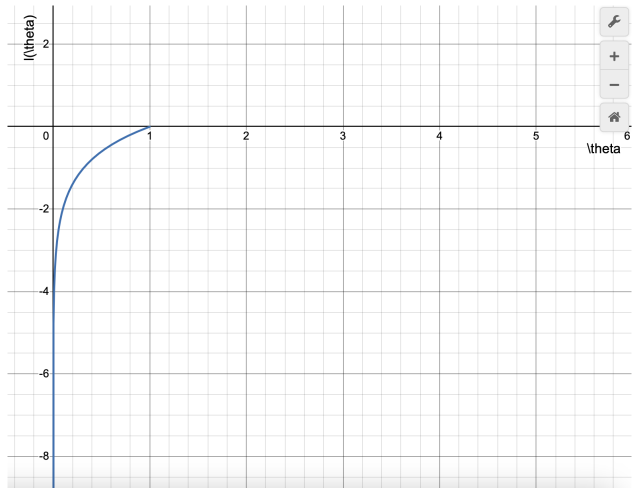

Figure 5: $l(\theta) = \log(\theta^2)$, $0 \leq \theta \leq 1$

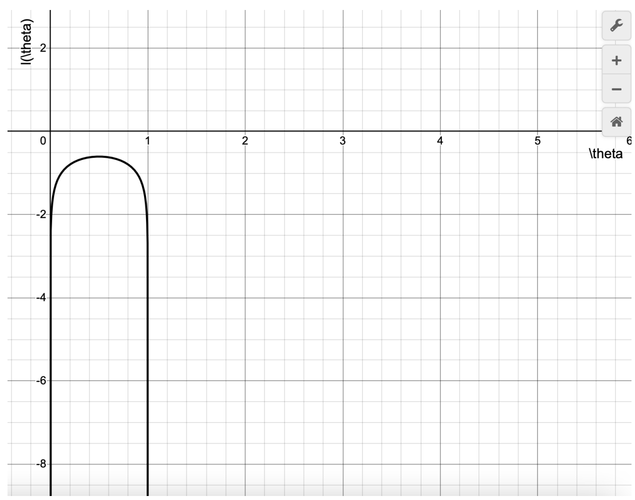

Figure 6: $l(\theta) = \log(\theta(1-\theta))$, $0 \leq \theta \leq 1$

It is reasonable to rescale $L(\theta)$ to have maximum value 1 and recenter $l(\theta)$ to have maximum value 0 because this makes comparison easier and makes the pattern more observable, including helping distinguish the maximum from the rest when we have a vector of numbers extremely close to 0. On the other hand, multiplicative constants become additive constants so recentering does not change the shape/curvature of the log.

We now move on to the discussion of **sufficient statistics** and the **factorization theorem**.

Definition: A statistic $T(X_1, \ldots X_n)$ is said to be **sufficient** for θ if the conditional distribution of $X_1, \ldots X_n$, given $T = t$, does not depend on θ for any value of t.

Theorem (Factorization Theorem): A summary statistic $T(X) = T(X_1, \ldots, X_n)$ is sufficient for a parameter θ if and only if $L(\theta)$ factors in the form $L(\theta) = g(T(X), \theta)h(x)$.
A sufficient statistic is a data summary that allows us to graph the likelihood function, scaled to

have a maximum of 1, or the log-likelihood function, recentered to have a maximum of 0. This theorem essentially helps us identify the "crucial part" in graphing (i.e. what the data actually depends on) and recentering/rescaling allows us to ignore the multiplicative/additive part in the equation.

Consider the following as an example, for the $n = 272$ games in the 2024 NFL season, the average winning margin for the home team was $\bar{x} = 1.95$. Assume that the winning margins are approximately iid $N(\theta, \sigma^2)$ random variables with known standard deviation $\sigma = 14$, we want to graph the likelihood function of θ.

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x_i - \theta)^2}{2\sigma^2}) \\
&\propto \exp(-\sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2\sigma^2}) \\
&= \exp(-\sum_{i=1}^{n} \frac{(x_i - \bar{x} + \bar{x} - \theta)^2}{2\sigma^2}) \\
&= \exp(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - \theta) + (\bar{x} - \theta)^2) \\
&= \exp((-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 + (\bar{x} - \theta)^2) \\
&\propto \exp(-\frac{1}{2\sigma^2} n(\bar{x} - \theta)^2) \\
&= \exp(-\frac{1}{2 \cdot 14^2} 272 \cdot (1.95 - \theta)^2)
\end{aligned}
$$

We will graph the last expression.

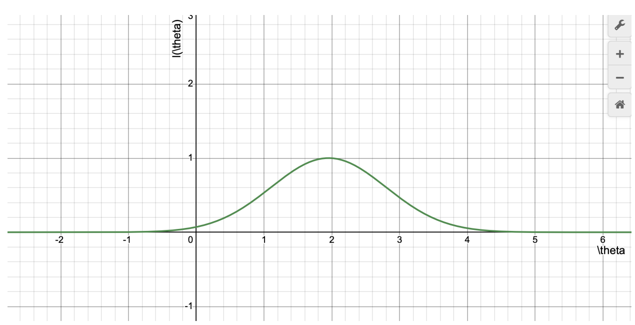

Figure 7: Graph of $L(\theta) = \exp(-\frac{1}{2 \cdot 14^2} 272 \cdot (1.95 - \theta)^2)$

Finally, we will define the Exponential family of distributions, and show how the k-parameter family includes only distributions for which there is a sufficient statistic of dimension k.

Definition (One-Parameter Exponential family): One-parameter members of the exponential family have density or frequency functions of the form

$$
\begin{aligned}
f(x \mid \theta) &= \exp[c(\theta)T(x) + d(\theta) + S(x)], \quad x \in A \\
&= 0, \quad x \notin A
\end{aligned}
$$

where the set A does not depend on θ.
k-parameter: A k-parameter member of the exponential family has a density or frequency function of the form

$$\begin{aligned} f(x \mid \theta) &= \exp[\sum_{i=1}^{k} c_i(\theta)T_i(x) + d(\theta) + S(x)], \quad x \in A \\ &= 0, \quad x \notin A \end{aligned}$$

where the set A does not depend on θ. Notice that in the formula we have a summation up to k, so this suggests that we don't need more than k parameters, which is essentially an upper bound on the number of parameters. (See below for how uniform fails to be included in the one-parameter exponential family). Therefore, the k-parameter family includes only distributions for which there is a sufficient statistic of dimension k.

Finally, we want to show that $N(\mu, \sigma^2)$ is included in the Exponential family but $Unif(\theta - 1/2, \theta + 1/2)$ is not.

Proof.

$$\begin{aligned} f(x \mid \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{(x-\mu)^2}{2\sigma^2}] \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{x^2 - 2\mu x + \mu^2}{2\sigma^2}] \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2\sigma^2}(x^2 - 2\mu x + \mu^2)] \\ &= \exp[-\frac{1}{2\sigma^2}(x^2 - 2\mu x + \mu^2) - \frac{1}{2}(\log 2\pi\sigma^2)] \end{aligned}$$

Assume that variance is known, let $c(\mu, \sigma^2) = \frac{\mu}{\sigma^2}$, $T(x) = x$, $d(\mu, \sigma^2) = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}(\log 2\pi\sigma^2)$, and $S(x) = -\frac{x^2}{2\sigma^2}$ If both unknown, then this will follow a two-parameter family, with the form

$$f(x; \theta_1, \theta_2) = \exp[c_1(\theta_1, \theta_2)T_1(x) + c_2(\theta_1, \theta_2)T_2(x) + d(\theta_1, \theta_2) + S(x)], x \in A$$

Let $T_1(x) = x$, $T_2(x) = x^2$, $c_1(\theta_1, \theta_2) = \frac{\mu}{\sigma^2}$, $c_2(\theta_1, \theta_2) = -\frac{1}{2\sigma^2}$, $d(\theta_1, \theta_2) = -\frac{1}{2}(\frac{\mu^2}{\sigma^2} + \log 2\pi\sigma^2)$, and $S(x) = 0$.

□

Proof. Let's now show that $Unif(\theta - 1/2, \theta + 1/2)$ is not in the one-dimensional exponential family because we cannot identify a 1-dimensional sufficient statistics. But for distributions inside the 1-parameter exponential family, there is always a univariate sufficient statistic.
We can write the pdf of x_i as $f(x; \theta) = I(x_i > \theta - 1/2)I(x_i < \theta + 1/2)$. We know that the likelihood function will be a product of these and can be written as $L(\theta) = I(\max(x_i) < \theta + 1/2)I(\min(x_i) > \theta - 1/2) = I(\max(x_i) - 1/2 < \theta < \min(x_i) + 1/2)$. However, in this case, we need a 2-dimensional statistic $(\min(x_i), \max(x_i)$ for a univariate parameter.

□