Week 5: Maximum Likelihood Estimation

STA238: Probability, Statistics, and Data Analysis II

Luis Sierra Muntané

2026-02-03

Recap: CLT

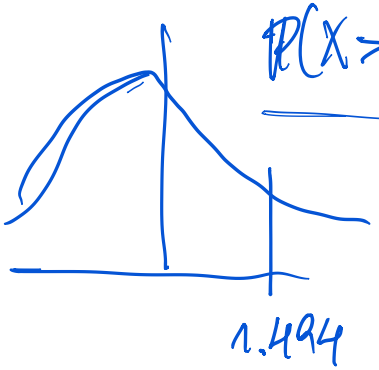
A high-frequency trading server in Toronto receives data packets according to a Poisson process. Based on historical data, the server receives an average of 5 packets per second. $X_i \sim Pois(5)$ → independent, $n > 30$ discrete

The system administrator is stress-testing the server's capacity. They want to calculate the probability that the server receives more than 18,200 packets in a single hour (3600 seconds).

$$X = \sum_{i=1}^{3600} X_i \sim Pois(5 \cdot 3600) = Pois(18000) \longrightarrow \mathbb{E}[X] = 18000, \sqrt{Var(X)} = \sqrt{18000}$$

By the CLT: $\frac{18200.5 - 18000}{134.16} \to N(0,1) \Longrightarrow \frac{200.5}{134.16} \approx 1.494$

$$\mathbb{P}(X > 18200) \approx \mathbb{P}(Z > 1.494) = 1 - \Phi(1.494) = 0.0676 \approx 7\%$$



Likelihood function

We commonly use the notation $f(x;\theta)$ to denote the density (or probability mass) at the point x when the underlying population parameters are θ.

If we have a random realization $x_1, \ldots, x_n$ and assume that it is drawn from a distribution with density (or PMF) f, we can think about how likely it would have been to observe that value: $f(x_i;\theta)$, where aggregating for the entire realization we would get

$$\mathcal{L}(x;\theta) := \prod_{i=1}^{n} f(x_i;\theta),$$

where we know think of the realizations as fixed, but θ as an unknown variable!

We call this the *Likelihood function*.

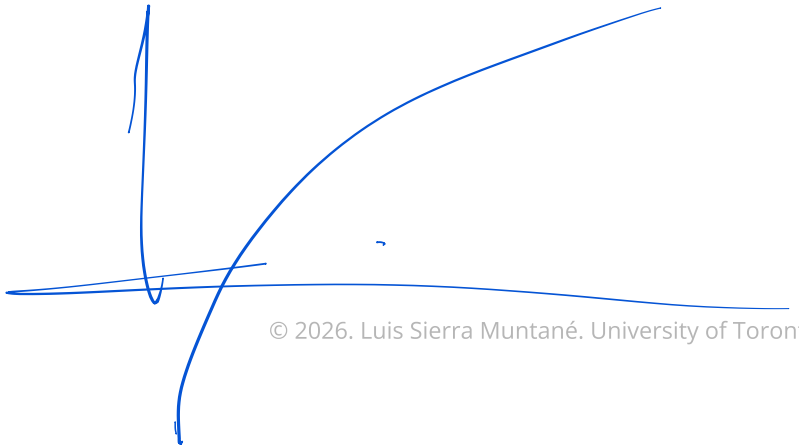
Likelihood function

Products are hard to work with, so we often prefer to analyze the *log*-likelihood

$$\ell(x;\theta) = \log \mathcal{L}(x;\theta) = \log \prod_{i=1}^{n} f(x_i;\theta) = \sum_{i=1}^{n} \log f(x_i;\theta),$$

where we use the fundamental property $\log(ab) = \log(a) + \log(b)$.

log is a monotone function $\Longrightarrow$ log is always increasing



Example: Coin Probability I

Suppose I have two unfair coins, one falls heads with probability $\theta_1 = 0.8$ and the other falls heads with probability $\theta_2 = 0.4$.

If I randomly pick a coin and it lands $\mathbf{TT}$, which is more *likely* to be the one I picked?

To compute the likelihood, note that $\mathbb{P}(\mathbf{T}) = (1-p)$, so by independence $\mathbb{P}(\mathbf{TT}) = (1-p)^2$.

If the coin is $\theta_1 = 0.8$, $\mathbb{P}(TT) = 0.2^2 = 0.04$

If the coin is $\theta_2 = 0.4$, $\mathbb{P}(TT) = 0.6^2 = 0.36$

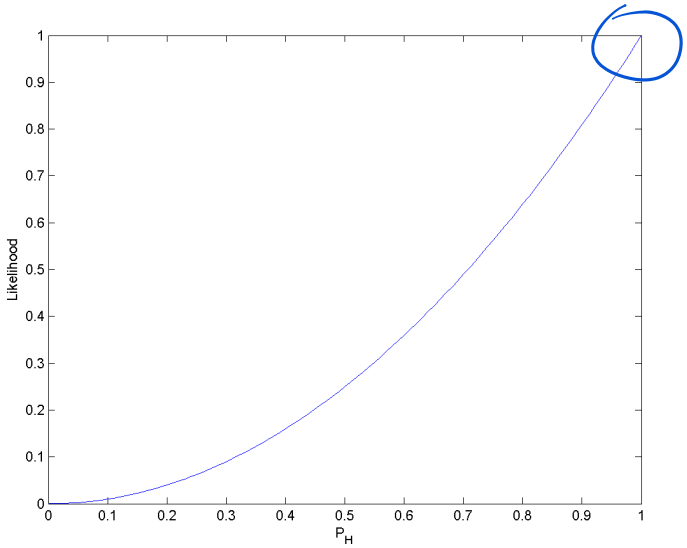
9:1 odds of the coin being θ_2

Example: Coin Probability II

Consider throwing an unfair coin two times and obtaining the sequence $\mathbf{HH}$.

To compute the likelihood, note that $\mathbb{P}(\mathbf{H}) = p$, so by independence $\mathbb{P}(\mathbf{HH}) = p^2$.

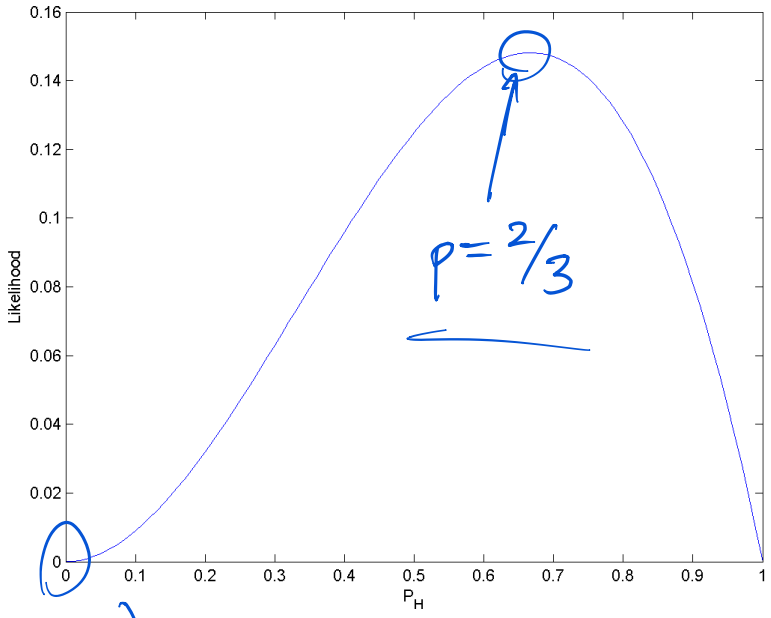
$\max_{p \in [0,1]} p^2 = 1$, $p = 1$



Example: Coin Probability II

Consider throwing an unfair coin three times and obtaining the sequence **HHT**.

What is the likelihood function in this case?



$$\mathcal{L}(x; p) = p \cdot p \cdot (1-p)$$

$$\max \; p^2(1-p) \;,$$

$$\frac{d}{dp}\left[p^2(1-p)\right] = p^2(-1) + 2p(1-p) = 0$$

$$-p^2 + 2p - 2p^2 = -3p^2 + 2p = p(2-3p) = 0$$

$$p = \frac{2}{3} \;, \quad \frac{d^2}{dp^2}\mathcal{L} = -2p - 2p + 2(1-p) = 2 - 6p$$

$$\left.\frac{d^2}{dp^2}\mathcal{L}\right|_{p=2/3} = 2 - \frac{2}{3}\cdot 6 = 2 - 4 = -2 < 0 \implies \text{local maximum.}$$

Example: Normal distribution

For a normal distribution, the likelihood of a random sample $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ can be found from the density of the normal distribution

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \implies \prod_{i=1}^{n} f(x_i; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

$$\mathcal{L} = (2\pi)^{-n/2} (\sigma^2)^{-n/2} \; e^{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2} e^{-\frac{1}{2}\left(\frac{x_2-\mu}{\sigma}\right)} \cdots$$

$$= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \; e^{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2 + \left(-\frac{1}{2}\left(\frac{x_2-\mu}{\sigma}\right)^2\right) + \cdots + \left(-\frac{1}{2}\left(\frac{x_n-\mu}{\sigma}\right)^2\right)}$$

$$= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \; e^{-\frac{1}{2}\left(\frac{1}{\sigma^2} \cdot \left[\sum_{i=1}^{n} (x_i-\mu)^2\right]\right)}$$

$$\ell(x; \mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2} \cdot \frac{1}{\sigma^2}\left[\sum_{i=1}^{n} (x_i-\mu)^2\right]$$

Maximum Likelihood Estimator (MLE)

Once we have our likelihood function, we can choose as our estimator the value that maximizes this likelihood

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \ell(\{X_i\}; \theta)$$

r.v. r.v.

And correspondingly, the estimate will be the value this estimator takes on for the specific realization $\{x_1, \ldots, x_n\}$.

Strategy when ℓ is a smooth function: solve

$$\nabla_\theta \ell(x; \theta) = \nabla_\theta \left[\sum_{i=1}^{n} \log f(x_i; \theta) \right] = \sum_{i=1}^{n} \nabla_\theta \log f(x_i; \theta) = 0.$$

$\frac{\partial}{\partial \theta}$

Example: Exponential distribution

Suppose we have independent realizations $x_1, \ldots, x_n$ from an Exponential $\mathbf{Exp}(\lambda)$ distribution.

$\lambda = \mathbb{E}\left[\frac{1}{X}\right]$

The density is given by $f(x_i; \lambda) = \lambda e^{-\lambda x_i}$, so the likelihood is given by:

$$\mathcal{L}(x; \lambda) = \prod_{i=1}^{n} f(x_i; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda\left(\sum_{i=1}^{n} x_i\right)}.$$

$$\max_{\lambda} \mathcal{L}(x; \lambda) = \max_{\lambda} \lambda^n e^{-\lambda\left(\sum_{i=1}^{n} x_i\right)}, \quad \log \mathcal{L}(x; \lambda) = n \log(\lambda) - \lambda\left(\sum_{i=1}^{n} x_i\right)$$

$$\frac{\partial}{\partial \lambda} \ell(x; \lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \implies \hat{\lambda}_{MLE} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}.$$

$\mathbb{E}[g(X)] \geq g(\mathbb{E}(X))$, $\frac{1}{x}$ is a strictly convex function $\implies$ Biased!

Example: Normally distributed samples

Suppose $x_1, \ldots, x_n$ are iid. realizations from $\mathcal{N}(\mu, \sigma^2)$.

We saw that the log-likelihood of a Gaussian was given by

$\frac{1}{2}\sigma^{-2} \longrightarrow -1\sigma^{-3}$

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

By differentiating and solving for μ, σ^2 we obtain $\hat{\mu}_{\text{MLE}} = \bar{x}_n$, $\hat{\sigma^2}_{\text{MLE}} = \frac{n-1}{n}S_n^2$:

$$\frac{\partial}{\partial\mu}\ell(\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)\cdot 2(-1) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\sum_{i=1}^{n}x_i - n\mu = 0 \longrightarrow \hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}x_i$$

$$\frac{\partial}{\partial\sigma^2}\ell(\mu,\sigma^2) = +\frac{n}{2}\cdot\frac{2}{\sigma} - \sigma^{-3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$+n\sigma^2 - \sum_{i=1}^{n}(x_i-\mu)^2 = 0 \longrightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{X})^2$$ → Biased

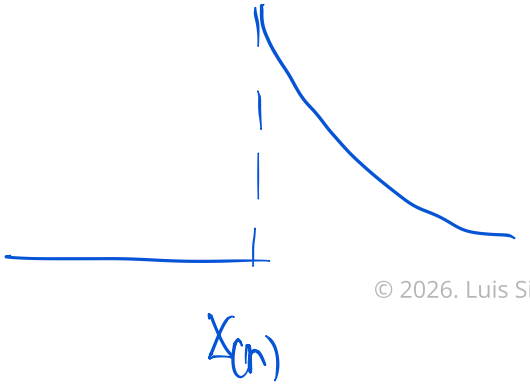
Example: Uniform distribution

Suppose $x_1, \ldots, x_n$ are iid. realizations now from $\mathbf{Unif}(0, \theta)$.

The density is $f(x_i; \theta) = \begin{cases} 1/\theta, & x_i \in [0, \theta] \\ 0, & x_i \notin [0, \theta] \end{cases} = \frac{1}{\theta} \mathbb{I}_{\{0 \leq x_i \leq \theta\}}$, so the likelihood is

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{I}_{\{0 \leq x_i \leq \theta\}} = \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbb{I}_{\{0 \leq x_i \leq \theta\}},$$

The product of the indicator functions $\prod_{i=1}^{n} \mathbb{I}_{\{0 \leq x_i \leq \theta\}}$ is equal to 1 if all x_i are between 0 and θ, and 0 otherwise.



$$\mathcal{L}(\theta) = \begin{cases} \frac{1}{\theta^n} & \text{if } \theta \geq X_{(n)} \\ 0 & \text{if } \theta < X_{(n)} \end{cases}.$$

$\max \mathcal{L}(\theta) \longrightarrow \hat{\theta} = X_{(n)}$

Biased! $\mathbb{E}[X_{(n)}] = \frac{n}{n+1}\theta$.

Distribution of scores*

For a random variable X_i, the quantity $\nabla_\theta \log f(X_i;\theta)$ or $\frac{\partial}{\partial\theta}\log f(X_i;\theta)$, known as the *score function*, is also a random variable. As such,

$$\ell(\{X_i\};\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta}\log f(X_i;\theta))$$

is a sum of iid. random variables. For this reason, we can apply the CLT to it.

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(X_i;\theta)\right] = \int_{-\infty}^{\infty} \frac{\frac{\partial}{\partial\theta} f(x_i;\theta)}{f(x_i;\theta)} \cdot f(x_i;\theta)\,dx = \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} f(x_i;\theta)\,dx = \frac{\partial}{\partial\theta} 1 = 0.$$

$$\mathbf{Var}\left[\frac{\partial}{\partial\theta}\log f(X_i;\theta)\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X_i;\theta)\right] =: I(\theta).$$

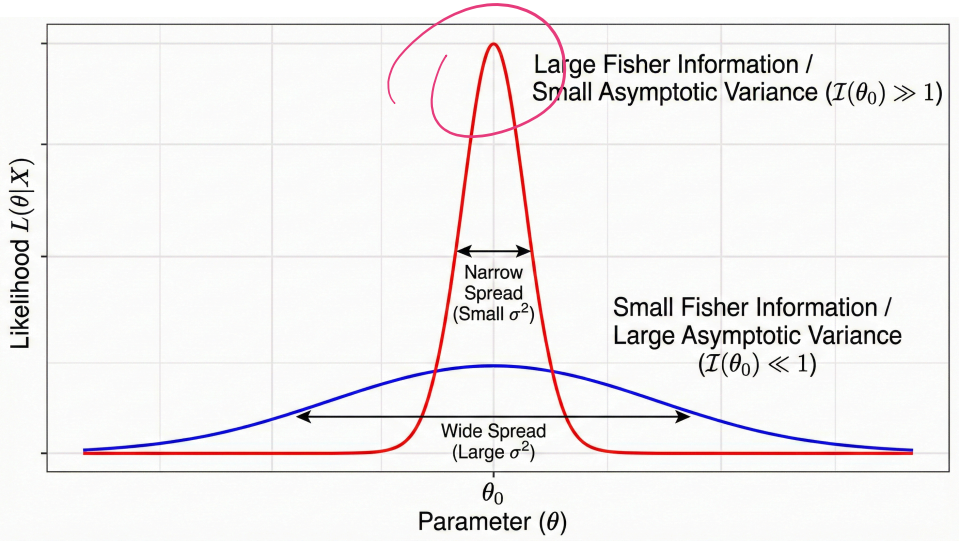
(integration by parts)

→ Fisher Information

Fisher Information

Some more intuition on the Fisher information from the relation

$$I(\theta) = \mathrm{Var}\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right].$$



Distribution of the MLE I

Suppose the true parameter value is θ_0. Let $S(\theta) = \frac{\partial}{\partial\theta}\log f(X;\theta)$, then we can use a Taylor expansion

true parameter value

$$S(\hat{\theta}) \approx S(\theta_0) + (\hat{\theta} - \theta_0)S'(\theta_0) \implies \hat{\theta} - \theta_0 = -\frac{S(\theta_0)}{S'(\theta_0)}.$$

- By the CLT, $\frac{1}{\sqrt{n}}S(\theta_0) \to \mathcal{N}(0, I(\theta_0))$
- By the WLLN, $\frac{1}{n}S'(\theta_0)$ converges to its expectation, which is $-I(\theta_0)$.

We can combine these facts to get

$$\sqrt{n}(\hat{\theta} - \theta_0) \to \frac{\mathcal{N}(0, I(\theta_0))}{I(\theta_0)}.$$

Distribution of the MLE II

As such, we have the score function is asymptotically normally distributed

$$\hat{\theta}_{\text{MLE}} \xrightarrow{d} \mathcal{N}\left(\theta_0, \frac{I(\theta_0)}{nI(\theta_0)^2}\right) = \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right).$$

> This "inverse" relationship says that the more information we have (higher $I(\theta)$), the more certain our estimate becomes (lower variance $I(\theta)^{-1}$).

Asymptotically normal

Asymptotically unbiased

$\Longrightarrow$ MLE is consistent

MLE Summary

From the previous slides, we can conclude that:

- The MLE is defined when the assumptions of the CLT hold for the score function
- The MLE is an *asymptotically* unbiased estimator, with expectation the true parameter value θ_0
- The MLE is a **consistent** estimator
- The MLE has asymptotic variance equal to the Fisher information $nI(\theta_0)$

Cramér-Rao Lower Bound

As it turns out, an asymptotic variance of $I(\theta)^{-1}$ is as good as we can do for our continuous, smooth estimators. For any distribution $X \sim f_\theta$ where θ represents the parameter we are estimating,

$$\mathbf{Var}(\hat{\theta}) \geq \frac{1}{n\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2\right]}$$

$I(\theta)$

MLE is asymptotically the most efficient estimator

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2]$$

$\mathbb{E}[X]^2 = 0$

Invariance of the MLE

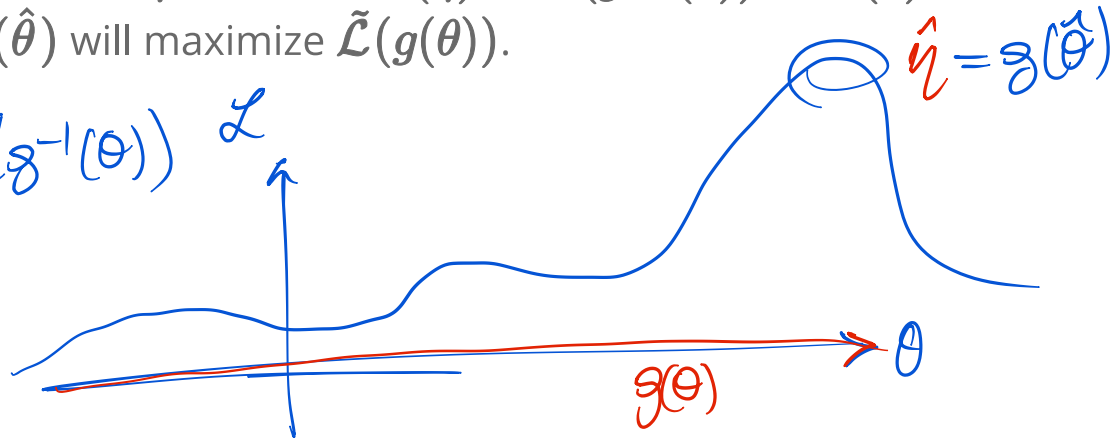
Recall that Jensen's inequality told us that if we had an unbiased estimator $T(X)$ for θ, then $g(T(X))$ would often be biased for $g(\theta)$.

Suppose $\hat{\theta}$ is the MLE of θ. For any function $g(\theta)$, the MLE of the transformation $\eta = g(\theta)$ is given by:

$$\hat{\eta} = g(\hat{\theta}).$$

> To maximize the likelihood $\tilde{\mathcal{L}}(\eta)$ in terms of η, note that $\tilde{\mathcal{L}}(\eta) = \mathcal{L}(g^{-1}(\theta)) = \mathcal{L}(\theta)$. If $\hat{\theta}$ maximizes the function $\mathcal{L}(\theta)$, then $g(\hat{\theta})$ will maximize $\tilde{\mathcal{L}}(g(\theta))$.

$$\underset{\eta}{\operatorname{argmax}}\ \tilde{\mathcal{L}}(\eta) = \underset{\theta}{\operatorname{argmax}}\ \mathcal{L}(g^{-1}(\theta))$$



Invariance of MLE: Example

We will now find the MLE for the standard deviation σ from normally distributed samples $\mathcal{N}(\mu, \sigma^2)$.

Recall that the MLE for the variance was $\hat{\sigma^2}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. We can obtain the MLE for the standard deviation by using the fact that $\sigma = \sqrt{\sigma^2}$, so

$$\arg\max_{\sigma>0} \tilde{\mathcal{L}}(\sigma) = \arg\max_{\sigma^2>0} \mathcal{L}(\sigma^2)$$

$\sigma = \sqrt{\sigma^2} \longrightarrow \hat{\sigma}_{MLE} = \sqrt{\hat{\sigma}^2_{MLE}}$

Small sample MLE

Going back to the coin example, image throwing an unfair coin two times and obtaining the sequence $\mathbf{HH}$.

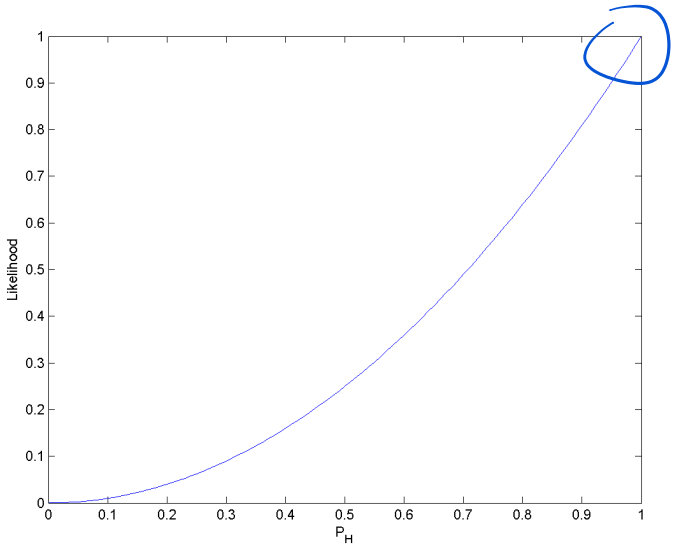
Data Sparsity

We saw the likelihood was given by $f(X_1, X_2; p) = \mathbb{P}(\mathbf{HH}) = p^2$.

The MLE is therefore given by

$$\hat{p}_{\mathrm{MLE}} = \arg\max_{p \in [0,1]} \{p^2\},$$

which leads us to conclude that the Maximum Likelihood Estimate for our realization $\mathbf{HH}$ is $\hat{p}_{\mathrm{MLE}} = 1$.



Assumptions:

In order to develop maximum likelihood estimation, we needed to make a few assumptions along the way:

- $X_1, \ldots, X_n$ is a random sample (iid.)
- $f(x_i; \theta)$ is not equal to zero (we don't observe any "impossible" points in our sample).
- The support of the distribution (the set of x values where $f(x; \theta) > 0$) does not depend on the parameter θ.
- Distinct parameter values ($\theta_1 \neq \theta_2$) must produce distinct probability distributions: $f(x; \theta_1) \neq f(x; \theta_2)$ (identifiable)
- The probability f is differentiable (or easy to maximize)

Recap Quiz

In which of the following scenarios would the CLT not be directly applicable to the sample mean?

a. Sampling from a Cauchy distribution. → $\mathbb{E}[X]$ and $\text{Var}(X)$ don't exist

b. Sampling from a Binomial distribution where $p = 0.01$.

c. Sampling from a Uniform distribution.

d. Sampling from a very small population without replacement. → could still be applied

Recap Quiz:

How is the Central Limit Theorem important for statistical tasks?

a. It eliminates the need for large sample sizes in experiments.

b. It allows us to use normal-based inference even when we don't know the population distribution.

c. It guarantees that the sample mean is always equal to the population mean.

d. It proves that all populations are eventually normal.

Recap Quiz:

Consider the iid. sample $X_1, \ldots, X_n$ and let $S_n = X_1 + \cdots + X_n$, $\bar{X}_n$ be the sample mean. Which of the following is an asymptotically standard normal random variable?

a. $\frac{\bar{X}_n - \mu}{\sigma^2}$

b. $\frac{S_n - n\mu}{n\sigma}$

c. $\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$

d. $\frac{S_n - \mu}{\sigma\sqrt{n}}$

"eventually" "for large enough n" $N(0,1)$

$N(\mu, \sigma^2)$

$\mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mu$, $\mathbb{E}[S_n] = \sum_{i=1}^{n}\mathbb{E}[X_i] = n\mu$

μ

$\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n} \longrightarrow \frac{\sigma}{\sqrt{n}}$

$\mathrm{Var}(S_n) = \sum_{i=1}^{n}\mathrm{Var}(X_i) = n\sigma^2$, $\longrightarrow \sqrt{n}\,\sigma$

$\frac{S_n - n\mu}{\sigma\sqrt{n}} \sim N(0,1)$ asymptotically.

Summary

- We have seen some uses for the CLT in statistics.
- Introduced the likelihood function and some of its characteristics.
- Developed and justified the method of Maximum Likelihood Estimation
- Showed and described the asymptotic properties of the MLE