- MOM estimators can be nonsensical. In fact, sometimes MOM estimators fall outside the parameter space Θ. For example, in linear models with random effects, variance components estimated via MOM can be negative.

7.2.2 Maximum likelihood estimation

Note: We first formally define a likelihood function; see also Section 6.3 (CB).

Definition: Suppose $\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$. Given that $\mathbf{X} = \mathbf{x}$ is observed, the function

$$L(\boldsymbol{\theta}|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$$

is called the **likelihood function**.

Note: The likelihood function $L(\boldsymbol{\theta}|\mathbf{x})$ is the same function as the joint pdf/pmf $f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$. The only difference is in how we interpret each one.

- The function $f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$ is a model that describes the random behavior of $\mathbf{X}$ when $\boldsymbol{\theta}$ is fixed.
- The function $L(\boldsymbol{\theta}|\mathbf{x})$ is viewed as a function of $\boldsymbol{\theta}$ with the data $\mathbf{X} = \mathbf{x}$ held fixed.

Interpretation: When $\mathbf{X}$ is discrete,

$$L(\boldsymbol{\theta}|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x}).$$

That is, when $\mathbf{X}$ is discrete, we can interpret the likelihood function $L(\boldsymbol{\theta}|\mathbf{x})$ literally as a joint probability.

- Suppose that $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are two possible values of $\boldsymbol{\theta}$. Suppose $\mathbf{X}$ is discrete and

$$L(\boldsymbol{\theta}_1|\mathbf{x}) = P_{\boldsymbol{\theta}_1}(\mathbf{X} = \mathbf{x}) > P_{\boldsymbol{\theta}_2}(\mathbf{X} = \mathbf{x}) = L(\boldsymbol{\theta}_2|\mathbf{x}).$$

  This suggests the sample $\mathbf{x}$ is more likely to have occurred with $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ rather than if $\boldsymbol{\theta} = \boldsymbol{\theta}_2$. Therefore, in the discrete case, we can interpret $L(\boldsymbol{\theta}|\mathbf{x})$ as "the probability of the data $\mathbf{x}$."

- Of course, this interpretation of $L(\boldsymbol{\theta}|\mathbf{x})$ is not appropriate when $\mathbf{X}$ is continuous because $P_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x}) = 0$. However, this description is still used informally when describing the likelihood function with continuous data. An attempt to make this description mathematical is given on pp 290 (CB).
- Section 6.3 (CB) describes how the likelihood function $L(\boldsymbol{\theta}|\mathbf{x})$ can be viewed as a **data reduction** device.

Definition: Any maximizer $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{x})$ of the likelihood function $L(\boldsymbol{\theta}|\mathbf{x})$ is called a **maximum likelihood estimate**.

- With our previous interpretation, we can think of $\widehat{\boldsymbol{\theta}}$ as "the value of $\boldsymbol{\theta}$ that maximizes the probability of the data $\mathbf{x}$."

We call $\widehat{\boldsymbol{\theta}}(\mathbf{X})$ a **maximum likelihood estimator** (MLE).

Remarks:

1. Finding the MLE $\widehat{\boldsymbol{\theta}}$ is essentially a maximization problem. The estimate $\widehat{\boldsymbol{\theta}}(\mathbf{x})$ must fall in the parameter space Θ because we are maximizing $L(\boldsymbol{\theta}|\mathbf{x})$ over Θ; i.e.,
$$\widehat{\boldsymbol{\theta}}(\mathbf{x}) = \arg\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}|\mathbf{x}).$$
There is no guarantee that an MLE $\widehat{\boldsymbol{\theta}}(\mathbf{x})$ will be unique (although it often is).
2. Under certain conditions (so-called "regularity conditions"), maximum likelihood estimators $\widehat{\boldsymbol{\theta}}(\mathbf{X})$ have very nice large-sample properties (Chapter 10, CB).
3. In most "real" problems, the likelihood function $L(\boldsymbol{\theta}|\mathbf{x})$ must be maximized numerically to calculate $\widehat{\boldsymbol{\theta}}(\mathbf{x})$.

Example 7.4. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{U}[0,\theta]$, where $\theta > 0$. Find the MLE of θ.
Solution. The likelihood function is
$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\theta} I(0 \le x_i \le \theta) = \underbrace{\frac{1}{\theta^n} I(x_{(n)} \le \theta) \prod_{i=1}^{n} I(x_i \ge 0)}_{\text{view this as a function of } \theta \text{ with } \mathbf{x} \text{ fixed}}.$$

$\mathcal{U}(0,\theta)$

$L(\theta|x) = \frac{1}{\theta^n} I(x_{(n)} < \theta) \prod_{i=1}^{n} I(x_i > 0)$

- $\theta > x_{(n)}$ $L(\theta|x) = \frac{1}{\theta^n}$
- $\theta \le x_{(n)}$ $L(\theta|x) = 0$

Note that

- For $\theta \ge x_{(n)}$, $L(\theta|\mathbf{x}) = 1/\theta^n$, which decreases as θ increases.
- For $\theta < x_{(n)}$, $L(\theta|\mathbf{x}) = 0$.

Clearly, the MLE of θ is $\widehat{\theta} = X_{(n)}$.

Remark: Note that in this example, we "closed the endpoints" on the support of X; i.e., the pdf of X is
$$f_X(x|\theta) = \begin{cases} \frac{1}{\theta}, & 0 \le x \le \theta \\ 0, & \text{otherwise.} \end{cases}$$
Mathematically, this model is no different than had we "opened the endpoints." However, if we used open endpoints, note that
$$x_{(n)} < \arg\max_{\theta>0} L(\theta|\mathbf{x}) < x_{(n)} + \epsilon$$
for all $\epsilon > 0$, and therefore the maximizer of $L(\theta|\mathbf{x})$; i.e., the MLE, would not exist.

Curiosity: In this uniform example, we derived the MOM estimator to be $\widehat{\theta} = 2\overline{X}$ in Example 7.1. The MLE is $\widehat{\theta} = X_{(n)}$. Which estimator is "better?"

Note: In general, when the likelihood function $L(\boldsymbol{\theta}|\mathbf{x})$ is a differentiable function of $\boldsymbol{\theta}$, we can use calculus to maximize $L(\boldsymbol{\theta}|\mathbf{x})$. If an MLE $\widehat{\boldsymbol{\theta}}$ exists, it must satisfy

$$\frac{\partial}{\partial \theta_j} L(\widehat{\boldsymbol{\theta}}|\mathbf{x}) = 0, \quad j = 1, 2, ..., k.$$

Of course, second-order conditions must be verified to ensure that $\widehat{\boldsymbol{\theta}}$ is a maximizer (and not a minimizer or some other value).

Example 7.5. Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\theta, 1)$, where $-\infty < \theta < \infty$. The likelihood function is

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\theta)^2/2} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2}. \end{aligned}$$

The derivative

$$\begin{aligned} \frac{\partial}{\partial \theta} L(\theta|\mathbf{x}) &= \underbrace{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2}}_{\text{this can never be zero}} \sum_{i=1}^{n}(x_i - \theta) \stackrel{\text{set}}{=} 0 \\ \Longrightarrow & \sum_{i=1}^{n}(x_i - \theta) = 0. \end{aligned}$$

Therefore, $\widehat{\theta} = \overline{x}$ is a first-order critical point of $L(\theta|\mathbf{x})$. Is $\widehat{\theta} = \overline{x}$ a maximizer? I calculated

$$\frac{\partial^2}{\partial \theta^2} L(\theta|\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2} \left\{ \left[\sum_{i=1}^{n}(x_i - \theta)\right]^2 - n \right\}.$$

!!!

Because

$$\left.\frac{\partial^2}{\partial \theta^2} L(\theta|\mathbf{x})\right|_{\theta=\overline{x}} = -n \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2} < 0,$$

the function $L(\theta|\mathbf{x})$ is concave down when $\theta = \overline{x}$; i.e., $\widehat{\theta} = \overline{x}$ maximizes $L(\theta|\mathbf{x})$. Therefore,

$$\widehat{\theta} = \widehat{\theta}(\mathbf{X}) = \overline{X}$$

is the MLE of θ.

Illustration: Under the $\mathcal{N}(\theta, 1)$ model assumption, I graphed in Figure 7.1 the likelihood function $L(\theta|\mathbf{x})$ after observing $x_1 = 2.437$, $x_2 = 0.993$, $x_3 = 1.123$, $x_4 = 1.900$, and $x_5 = 3.794$ (an iid sample of size $n = 5$). The sample mean $\overline{x} = 2.049$ is our ML estimate of θ based on this sample $\mathbf{x}$.

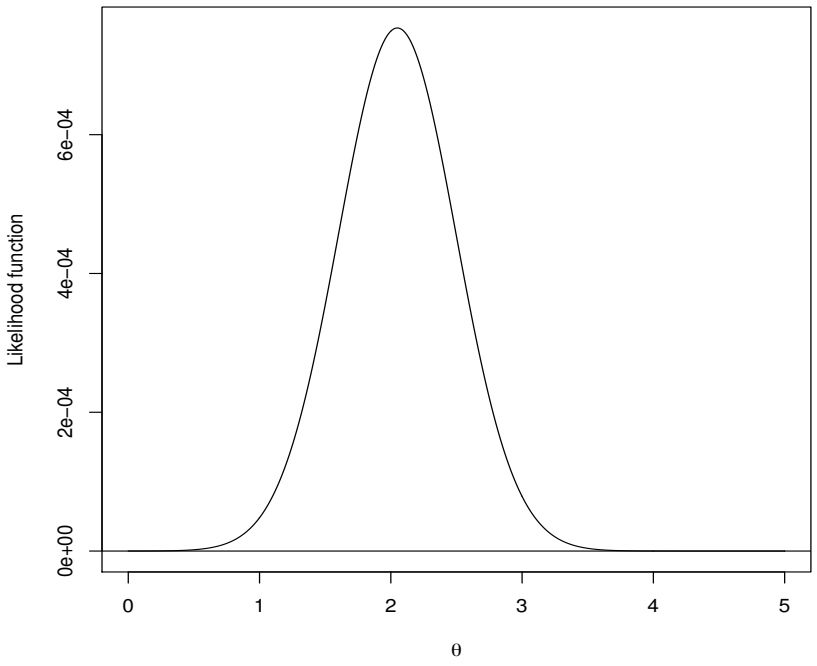

Figure 7.1: Plot of $L(\theta|\mathbf{x})$ versus θ in Example 7.5. The data $\mathbf{x}$ were generated from a $\mathcal{N}(\theta = 1.5, 1)$ distribution with $n = 5$. The sample mean (MLE) is $\overline{x} = 2.049$.

Q: What if, in Example 7.5, we constrained the parameter space to be $\Theta_0 = \{\theta : \theta \geq 0\}$? What is the MLE over Θ_0?
A: We simply maximize $L(\theta|\mathbf{x})$ over Θ_0 instead. It is easy to see the restricted MLE is

$$\widehat{\theta}^* = \widehat{\theta}^*(\mathbf{X}) = \begin{cases} \overline{X}, & \overline{X} \geq 0 \\ 0, & \overline{X} < 0. \end{cases}$$

Important: Suppose that $L(\boldsymbol{\theta}|\mathbf{x})$ is a likelihood function. Then

$$\begin{aligned} \widehat{\boldsymbol{\theta}}(\mathbf{x}) &= \arg\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}|\mathbf{x}) \\ &= \arg\max_{\boldsymbol{\theta}\in\Theta} \ln L(\boldsymbol{\theta}|\mathbf{x}). \end{aligned}$$

The function $\ln L(\boldsymbol{\theta}|\mathbf{x})$ is called the **log-likelihood function**. Analytically, it is usually easier to work with $\ln L(\boldsymbol{\theta}|\mathbf{x})$ than with the likelihood function directly. The equations

$$\frac{\partial}{\partial\theta_j} \ln L(\boldsymbol{\theta}|\mathbf{x}) = 0, \quad j = 1, 2, ..., k,$$

are called the **score equations**.

Example 7.6. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$; i.e., both parameters are unknown. Set $\boldsymbol{\theta} = (\mu, \sigma^2)$. The likelihood function is

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}. \end{aligned}$$

The log-likelihood function is

$$\ln L(\boldsymbol{\theta}|\mathbf{x}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2. = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

The score equations are

$$\begin{aligned} \frac{\partial}{\partial\mu}\ln L(\boldsymbol{\theta}|\mathbf{x}) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) \overset{\text{set}}{=} 0 \Rightarrow \widehat{\mu} = \overline{x} \\ \frac{\partial}{\partial\sigma^2}\ln L(\boldsymbol{\theta}|\mathbf{x}) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 \overset{\text{set}}{=} 0. \Rightarrow \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \end{aligned}$$

Clearly $\widehat{\mu} = \overline{x}$ solves the first equation; inserting $\widehat{\mu} = \overline{x}$ into the second equation and solving for σ^2 gives $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$. A first-order critical point is $(\overline{x}, n^{-1}\sum_{i=1}^{n}(x_i-\overline{x})^2)$.

Q: How can we verify this solution is a maximizer?
A: In general, for a k-dimensional maximization problem, we can calculate the Hessian matrix

$$\mathbf{H} = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\ln L(\boldsymbol{\theta}|\mathbf{x}) =$$

a $k \times k$ matrix of second-order partial derivatives, and show this matrix is **negative definite** when we evaluate it at the first-order critical point $\widehat{\boldsymbol{\theta}}$. This is a sufficient condition. Recall a $k \times k$ matrix $\mathbf{H}$ is negative definite if $\mathbf{a}'\mathbf{H}\mathbf{a} < 0$ for all $\mathbf{a} \in \mathbb{R}^k$, $\mathbf{a} \neq \mathbf{0}$.

if $\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{pmatrix}$

$$\begin{pmatrix} \frac{\partial^2}{\partial\theta_1^2}\ln L(\theta|x) & \frac{\partial^2}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2}{\partial\theta_1\partial\theta_k} \\ \frac{\partial^2}{\partial\theta_2\partial\theta_1} & \frac{\partial^2}{\partial\theta_2^2} & \cdots & \frac{\partial^2}{\partial\theta_2\partial\theta_k} \\ \vdots & & & \\ \frac{\partial^2}{\partial\theta_k\partial\theta_1} & & \cdots & \frac{\partial^2}{\partial\theta_k^2} \end{pmatrix}$$

For the $\mathcal{N}(\mu, \sigma^2)$ example, I calculated

$$\mathbf{H} = \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu) \\ -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu) & \frac{n}{\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\mu)^2 \end{pmatrix}$$

With $\mathbf{a}' = (a_1, a_2)$, it follows that

$$\mathbf{a}'\mathbf{H}\mathbf{a}\Big|_{\mu=\widehat{\mu},\sigma^2=\widehat{\sigma}^2} = -\frac{na_1^2}{\widehat{\sigma}^2} < 0.$$

This shows that

$$\widehat{\boldsymbol{\theta}}(\mathbf{X}) = \begin{pmatrix} \overline{X} \\ \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2 \end{pmatrix}$$

is the MLE of $\boldsymbol{\theta}$ in the $\mathcal{N}(\mu, \sigma^2)$ model.

$\widehat{\mu}_{MLE} = \overline{X}_n$

$\widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2$

$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2$

Exercise: Find the MLEs of μ and σ^2 in the respective sub-families:

- $\mathcal{N}(\mu, \sigma_0^2)$, where σ_0^2 is known
- $\mathcal{N}(\mu_0, \sigma^2)$, where μ_0 is known.

Example 7.7. *ML estimation under parameter constraints.* Suppose X_1, X_2 are independent random variables where

$$\begin{aligned} X_1 &\sim b(n_1, p_1) \\ X_2 &\sim b(n_2, p_2), \end{aligned}$$

where $0 < p_1 < 1$ and $0 < p_2 < 1$. The likelihood function of $\boldsymbol{\theta} = (p_1, p_2)$ is

$$\begin{aligned} L(\boldsymbol{\theta}|x_1, x_2) &= f_{X_1}(x_1|p_1) f_{X_2}(x_2|p_2) \\ &= \binom{n_1}{x_1} p_1^{x_1}(1-p_1)^{n_1-x_1} \binom{n_2}{x_2} p_2^{x_2}(1-p_2)^{n_2-x_2}. \end{aligned}$$

The log-likelihood function is

$$\ln L(\boldsymbol{\theta}|x_1, x_2) = c + x_1 \ln p_1 + (n_1 - x_1)\ln(1-p_1) + x_2 \ln p_2 + (n_2 - x_2)\ln(1-p_2),$$

where $c = \ln \binom{n_1}{x_1} + \ln \binom{n_2}{x_2}$ is free of $\boldsymbol{\theta}$. Over

$$\Theta = \{\boldsymbol{\theta} = (p_1, p_2) : 0 < p_1 < 1, 0 < p_2 < 1\},$$

it is easy to show that $\ln L(\boldsymbol{\theta}|x_1, x_2)$ is maximized at

$$\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(X_1, X_2) = \begin{pmatrix} \widehat{p}_1 \\ \widehat{p}_2 \end{pmatrix} = \begin{pmatrix} \dfrac{X_1}{n_1} \\ \dfrac{X_2}{n_2} \end{pmatrix},$$

the vector of sample proportions. Because this is the maximizer over the entire parameter space Θ, we call $\widehat{\boldsymbol{\theta}}$ the **unrestricted MLE** of $\boldsymbol{\theta}$.

Q: How do we find the MLE of $\boldsymbol{\theta}$ subject to the constraint that $p_1 = p_2$?
A: We would now like to maximize $\ln L(\boldsymbol{\theta}|x_1, x_2)$ over

$$\Theta_0 = \{\boldsymbol{\theta} = (p_1, p_2) : 0 < p_1 < 1, 0 < p_2 < 1,\ p_1 = p_2\}.$$

We can use Lagrange multipliers to maximize $\ln L(\boldsymbol{\theta}|x_1, x_2)$ subject to the constraint that

$$g(\boldsymbol{\theta}) = g(p_1, p_2) = p_1 - p_2 = 0.$$

We are left to solve

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}} \ln L(\boldsymbol{\theta}|x_1, x_2) &= \lambda \frac{\partial}{\partial \boldsymbol{\theta}} g(\boldsymbol{\theta}) \\ g(\boldsymbol{\theta}) &= 0. \end{aligned}$$

This system becomes

$$\begin{aligned} \frac{x_1}{p_1} - \frac{n_1 - x_1}{1 - p_1} &= \lambda \\ \frac{x_2}{p_2} - \frac{n_2 - x_2}{1 - p_2} &= -\lambda \\ p_1 - p_2 &= 0. \end{aligned}$$

Solving this system for p_1 and p_2, we get

$$\widehat{\boldsymbol{\theta}}^* = \widehat{\boldsymbol{\theta}}^*(X_1, X_2) = \begin{pmatrix} \widehat{p}_1^* \\ \widehat{p}_2^* \end{pmatrix} = \begin{pmatrix} \dfrac{X_1 + X_2}{n_1 + n_2} \\ \dfrac{X_1 + X_2}{n_1 + n_2} \end{pmatrix}.$$

Because this is the maximizer over the subspace Θ_0, we call $\widehat{\boldsymbol{\theta}}^*$ the **restricted MLE**; i.e., the MLE of $\boldsymbol{\theta}$ subject to the $p_1 = p_2$ restriction.

Discussion: The parameter constraint $p_1 = p_2$ might arise in a **hypothesis test**; e.g., $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$. If H_0 is true, then we would expect $\widehat{\boldsymbol{\theta}}^*$ and $\widehat{\boldsymbol{\theta}}$ to be "close" and the ratio

$$\lambda(x_1, x_2) = \frac{L(\widehat{\boldsymbol{\theta}}^*|x_1, x_2)}{L(\widehat{\boldsymbol{\theta}}|x_1, x_2)} \approx 1.$$

The farther $\widehat{\boldsymbol{\theta}}^*$ is from $\widehat{\boldsymbol{\theta}}$, the smaller $\lambda(x_1, x_2)$ becomes. Therefore, it would make sense to reject H_0 when $\lambda(x_1, x_2)$ is small. This is the idea behind **likelihood ratio tests**.

Example 7.8. *Logistic regression.* In practice, finding maximum likelihood estimates usually requires numerical methods. Suppose $Y_1, Y_2, ..., Y_n$ are independent Bernoulli random variables; specifically, $Y_i \sim \text{Bernoulli}(p_i)$, where

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i \iff p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}.$$

In this model, the x_i's are fixed constants. The likelihood function of $\boldsymbol{\theta} = (\beta_0, \beta_1)$ is

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{y}) &= \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} \\ &= \prod_{i=1}^{n} \left[\frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}\right]^{y_i} \left[1 - \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}\right]^{1-y_i}. \end{aligned}$$

Taking logarithms and simplifying gives

$$\ln L(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} \left[y_i(\beta_0 + \beta_1 x_i) - \ln(1 + e^{\beta_0 + \beta_1 x_i})\right].$$

$\frac{\partial \ln L(\theta|y)}{\partial \beta_0} = 0$

$\frac{\partial}{\partial \beta_1} = 0$

Closed-form expressions for the maximizers $\widehat{\beta}_0$ and $\widehat{\beta}_1$ do not exist except in very simple situations. Numerical methods are needed to maximize $\ln L(\boldsymbol{\theta}|\mathbf{y})$; e.g., iteratively re-weighted least squares (the default method in R's `glm` function).

Theorem 7.2.10 (Invariance property of MLEs). Suppose $\widehat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$. For any function $\tau(\boldsymbol{\theta})$, the MLE of $\tau(\boldsymbol{\theta})$ is $\tau(\widehat{\boldsymbol{\theta}})$.

Proof. For simplicity, suppose θ is a scalar parameter and that $\tau : \mathbb{R} \to \mathbb{R}$ is one-to-one (over Θ). In this case,

$$\eta = \tau(\theta) \iff \theta = \tau^{-1}(\eta).$$

The likelihood function of interest is $L^*(\eta)$. It suffices to show that $L^*(\eta)$ is maximized when $\eta = \tau(\widehat{\theta})$, where $\widehat{\theta}$ is the maximizer of $L(\theta)$. For simplicity in notation, I drop emphasis of a likelihood function's dependence on $\mathbf{x}$. Let $\widehat{\eta}$ be a maximizer of $L^*(\eta)$. Then

$$\begin{aligned} L^*(\widehat{\eta}) &= \sup_{\eta} L^*(\eta) \\ &= \sup_{\eta} L(\tau^{-1}(\eta)) \\ &= \sup_{\theta} L(\theta). \end{aligned}$$

Therefore, the maximizer $\widehat{\eta}$ satisfies $\tau^{-1}(\widehat{\eta}) = \widehat{\theta}$. Because τ is one-to-one, $\widehat{\eta} = \tau(\widehat{\theta})$. □

Remark: Our proof assumes that τ is a one-to-one function. However, Theorem 7.2.10 is true for any function; see pp 319-320 (CB).

Example 7.8 (continued). In the logistic regression model

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i \iff p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = \tau(\beta_0, \beta_1), \text{ say,}$$

the MLE of p_i is

$$\widehat{p}_i = \tau(\widehat{\beta}_0, \widehat{\beta}_1) = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_i)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_i)}.$$

Example 7.9. Suppose $X_1, X_2, ..., X_n$ are iid exponential(β), where $\beta > 0$. The likelihood function is

$$L(\beta|\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-\sum_{i=1}^n x_i/\beta}.$$

The log-likelihood function is

$$\ln L(\beta|\mathbf{x}) = -n \ln \beta - \frac{\sum_{i=1}^n x_i}{\beta}$$

The score equation becomes

$$\frac{\partial}{\partial \beta} \ln L(\beta|\mathbf{x}) = -\frac{n}{\beta} + \frac{\sum_{i=1}^n x_i}{\beta^2} \stackrel{\text{set}}{=} 0.$$

Solving the score equation for β gives $\widehat{\beta} = \overline{x}$. It is easy to show that this value maximizes $\ln L(\beta|\mathbf{x})$. Therefore,

$$\widehat{\beta} = \widehat{\beta}(\mathbf{X}) = \overline{X}$$

is the MLE of β.

Applications of invariance: In Example 7.9,

- $\overline{X}^2$ is the MLE of β^2
- $1/\overline{X}$ is the MLE of $1/\beta$
- For t fixed, $e^{-t/\overline{X}}$ is the MLE of $S_X(t|\beta) = e^{-t/\beta}$, the **survivor function** of X at t.

7.2.3 Bayesian estimation

Remark: Non-Bayesians think of inference in the following way:

$$\text{Observe } \mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x}|\theta) \quad \longrightarrow \quad \text{Use } \mathbf{x} \text{ to make statement about } \theta.$$

In this paradigm, the model parameter θ is fixed (and unknown). I have taken θ to be a scalar here for ease of exposition.

Bayesians do not consider the parameter θ to be fixed. They regard θ as random, having its own probability distribution. Therefore, Bayesians think of inference in this way:

$$\text{Model } \theta \sim \pi(\theta) \quad \longrightarrow \quad \text{Observe } \mathbf{X}|\theta \sim f_{\mathbf{X}}(\mathbf{x}|\theta) \quad \longrightarrow \quad \text{Update with } \pi(\theta|\mathbf{x}).$$

The model for θ on the front end is called the **prior distribution**. The model on the back end is called the **posterior distribution**. The posterior distribution **combines** prior information (supplied through the prior model) and the observed data $\mathbf{x}$. For a Bayesian, all inference flows from the posterior distribution.

Important: Here are the relevant probability distributions that arise in a Bayesian context. These are given "in order" as to how the Bayesian uses them. Continue to assume that θ is a scalar.

1. **Prior distribution:** $\theta \sim \pi(\theta)$. This distribution incorporates the information available about θ before any data are observed.
2. **Conditional distribution:** $\mathbf{X}|\theta \sim f_{\mathbf{X}}(\mathbf{x}|\theta)$. This is the distribution of $\mathbf{X}$, but now viewed conditionally on θ:

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\theta) &= L(\theta|\mathbf{x}) \\ &\overset{\text{iid}}{=} \prod_{i=1}^{n} f_{X|\theta}(x_i|\theta). \end{aligned}$$

   Mathematically, the conditional distribution is the same as the likelihood function.

3. **Joint distribution:** This distribution describes how $\mathbf{X}$ and θ vary jointly. From the definition of a conditional distribution,

$$f_{\mathbf{X},\theta}(\mathbf{x},\theta) = \underbrace{f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)}_{\text{likelihood}} \; \underbrace{\pi(\theta)}_{\text{prior}}.$$

4. **Marginal distribution.** This describes how $\mathbf{X}$ is distributed marginally. From the definition of a marginal distribution,

$$\begin{aligned} m_{\mathbf{X}}(\mathbf{x}) &= \int_{\Theta} f_{\mathbf{X},\theta}(\mathbf{x},\theta)d\theta \\ &= \int_{\Theta} f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta)d\theta, \end{aligned}$$

where Θ is the "support" of θ (remember, we are now treating θ as a random variable).

5. **Posterior distribution.** This is the Bayesian's "updated" distribution of θ, given that the data $\mathbf{X} = \mathbf{x}$ have been observed. From the definition of a conditional distribution,

$$\begin{aligned} \pi(\theta|\mathbf{x}) &= \frac{f_{\mathbf{X},\theta}(\mathbf{x},\theta)}{m_{\mathbf{X}}(\mathbf{x})} \\ &= \frac{f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta)}{\int_{\Theta} f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta)d\theta}. \end{aligned}$$

Remark: The process of starting with $\pi(\theta)$ and performing the necessary calculations to end up with $\pi(\theta|\mathbf{x})$ is informally known as "turning the Bayesian crank." The distributions above can be viewed as steps in a "recipe" for posterior construction (i.e., start with the prior and the conditional, calculate the joint, calculate the marginal, calculate the posterior). We will see momentarily that not all steps are needed. In fact, in practice, computational techniques are used to essentially bypass Step 4 altogether. You can see that this might be desirable, especially if θ is a vector (and perhaps high-dimensional).

Example 7.10. Suppose that, conditional on θ, $X_1, X_2, ..., X_n$ are iid Poisson(θ), where the prior distribution for $\theta \sim$ gamma(a, b), a, b known. We now turn the Bayesian crank.

1. **Prior distribution.**

$$\pi(\theta) = \frac{1}{\Gamma(a)b^a}\ \theta^{a-1}e^{-\theta/b}I(\theta > 0).$$

2. **Conditional distribution.** For $x_i = 0, 1, 2, ...$,

$$f_{\mathbf{X}|\theta}(\mathbf{x}|\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i}e^{-\theta}}{x_i!} = \frac{\theta^{\sum_{i=1}^n x_i}e^{-n\theta}}{\prod_{i=1}^n x_i!}.$$

Recall that this is the same function as the likelihood function.

3. **Joint distribution.** For $x_i = 0, 1, 2, ...$, and $\theta > 0$,

$$\begin{aligned} f_{\mathbf{X},\theta}(\mathbf{x},\theta) &= f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta) \\ &= \frac{\theta^{\sum_{i=1}^n x_i}e^{-n\theta}}{\prod_{i=1}^n x_i!}\ \frac{1}{\Gamma(a)b^a}\ \theta^{a-1}e^{-\theta/b} \\ &= \underbrace{\frac{1}{\prod_{i=1}^n x_i!\ \Gamma(a)b^a}}_{\text{does not depend on } \theta}\ \theta^{\sum_{i=1}^n x_i + a - 1}e^{-\theta/\left(n+\frac{1}{b}\right)^{-1}}. \end{aligned}$$

4. **Marginal distribution.** For $x_i = 0, 1, 2, ...,$

$$\begin{aligned} m_{\mathbf{X}}(\mathbf{x}) &= \int_{\Theta} f_{\mathbf{X},\theta}(\mathbf{x}, \theta) d\theta \\ &= \frac{1}{\prod_{i=1}^{n} x_i!\ \Gamma(a) b^a} \int_0^{\infty} \underbrace{\theta^{\sum_{i=1}^{n} x_i + a - 1} e^{-\theta/\left(n+\frac{1}{b}\right)^{-1}}}_{\text{gamma}(a^*, b^*)\text{ kernel}} d\theta, \end{aligned}$$

where

$$a^* = \sum_{i=1}^{n} x_i + a \quad \text{and} \quad b^* = \frac{1}{n + \frac{1}{b}}.$$

Therefore,

$$m_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\prod_{i=1}^{n} x_i!\ \Gamma(a) b^a}\ \Gamma\left(\sum_{i=1}^{n} x_i + a\right) \left(\frac{1}{n + \frac{1}{b}}\right)^{\sum_{i=1}^{n} x_i + a}.$$

5. **Posterior distribution.** For $\theta > 0$,

$$\begin{aligned} \pi(\theta|\mathbf{x}) &= \frac{f_{\mathbf{X},\theta}(\mathbf{x}, \theta)}{m_{\mathbf{X}}(\mathbf{x})} \\ &= \frac{\frac{1}{\prod_{i=1}^{n} x_i!\ \Gamma(a) b^a}\ \theta^{\sum_{i=1}^{n} x_i + a - 1} e^{-\theta/\left(n+\frac{1}{b}\right)^{-1}}}{\frac{1}{\prod_{i=1}^{n} x_i!\ \Gamma(a) b^a}\ \Gamma\left(\sum_{i=1}^{n} x_i + a\right) \left(\frac{1}{n + \frac{1}{b}}\right)^{\sum_{i=1}^{n} x_i + a}} \\ &= \frac{1}{\Gamma\left(\sum_{i=1}^{n} x_i + a\right) \left(\frac{1}{n+\frac{1}{b}}\right)^{\sum_{i=1}^{n} x_i + a}}\ \theta^{\sum_{i=1}^{n} x_i + a - 1} e^{-\theta/\left(n+\frac{1}{b}\right)^{-1}}, \end{aligned}$$

which we recognize as the gamma pdf with parameters

$$\begin{aligned} a^* &= \sum_{i=1}^{n} x_i + a \\ b^* &= \frac{1}{n + \frac{1}{b}}. \end{aligned}$$

That is, the posterior distribution

$$\theta|\mathbf{X} = \mathbf{x} \sim \text{gamma}\left(\sum_{i=1}^{n} x_i + a,\ \frac{1}{n + \frac{1}{b}}\right).$$

Remark: Note that the shape and scale parameters of the posterior distribution $\pi(\theta|\mathbf{x})$ depend on

- a and b, the prior distribution parameters (i.e., the "hyperparameters")
- the data $\mathbf{x}$ through the sufficient statistic $t(\mathbf{x}) = \sum_{i=1}^{n} x_i$.

In this sense, the posterior distribution combines information from the prior and the data.