Estimation methods

Contents

1 The method of moments (MoM) **2**

2 The method of maximum likelihood (ML) **5**

2.1 Likelihood: definition and meaning 5

2.2 MLE implementation 8

In this section, we examine some general methods that can be used to construct estimators that often have good properties.

1 The method of moments (MoM)

Let X_i, $i = 1, \ldots, n$, be an iid sample of X. Let $\mu_j = E(X^j)$ denote the $j-$th moment of X, and $\hat{\mu}_j = n^{-1} \sum_{i=1}^{n} X_i^j$ the $j-$th sample moment.
To apply the method of moments to the problem of estimating a parameter $\boldsymbol{\theta}$, we need to be able to express $\boldsymbol{\theta}$ as a function of $\mu_1, \mu_2, \ldots$. Thus, we need to find a (known) function, say $\boldsymbol{g}$, such that

$$\boldsymbol{\theta} = \boldsymbol{g}(\mu_1, \mu_2, \ldots).$$

A simple estimation method consists in replacing the μ_j's in the equation above by their empirical versions. This leads to the following MoM estimator:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{g}(\hat{\mu}_1, \hat{\mu}_2, \ldots).$$

Broadly speaking, MoM estimators do not necessarily offer the best performances, but they are typically easy to obtain and, under reasonable conditions, they are consistent and are asymptotically normal. In fact,

- By the WLLN + CMT, a MoM estimator is consistent provided (i) the population moments exist, and (ii) g is a continuous.
- By the CLT + Delta method, a MoM estimator can, typically, be shown to be asymptotically normal.

Example 1.1.

- Let X_i, $i = 1, \ldots, n$, be an iid sample from the exponential distribution with pd $f(x; \lambda) = \lambda e^{-\lambda x} I(x \geq 0)$, $\lambda > 0$. Since $E(X_1) = 1/\lambda$, the MoM estimator of λ is $\hat{\lambda} = 1/\overline{X}_n$. This is a consistent estimator. Moreover, since $Var(X_1) = 1/\lambda^2$, $\hat{\lambda}$ is asymptotically normal with limiting distribution given by $N(\lambda, \lambda^2/n)$.

- Let X_i, $i = 1, \ldots, n$, be an iid sample of X. Let $\mu_j = E(X^j)$, $\sigma_{jk} = Cov(X^j, X^k)$, $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$, and $\overline{X_n^2} = n^{-1} \sum_{i=1}^n X_i^2$. Let's find the MoM estimator of $\sigma^2 \equiv \sigma_{11} = Var(X)$ and its asymptotic distribution. Since $\sigma^2 = \mu_2 - \mu_1^2$, the MoM estimator of σ^2 is

$$\hat{\sigma}_n^2 = \hat{\mu}_2 - \hat{\mu}_1^2 = \overline{X_n^2} - \overline{X}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

By CLT,

$$\sqrt{n} \left(\begin{pmatrix} \overline{X}_n \\ \overline{X_n^2} \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \right) \xrightarrow{d} N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \right)$$

So, by the Delta method, with $g(x,y) = y - x^2$,

$$\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \xrightarrow{d} N(0, \nu^2),$$

where $\nu^2 := 4\mu_1^2\sigma_{11} + \sigma_{22} - 4\mu_1\sigma_{12} = Var(X - \mu_1)^2$.

- Let Y_i, $i = 1, \ldots, n$, be an iid sample from the Log-normal distribution (i.e. $log(Y) \sim N(\mu, \sigma^2)$) with parameters μ and σ^2. Y_i has the pd

$$f(y; \mu, \sigma^2) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log y - \mu)^2}{2\sigma^2}\right) I(y > 0),\ \mu \in (-\infty, \infty),\ \text{and } \sigma > 0.$$

Since

$$\mu_1 := E(Y_1) = \exp\left(\mu + \sigma^2/2\right) \text{ and } \mu_2 := E(Y_1^2) = \exp\left(2\mu + 2\sigma^2\right),$$

the MoM estimator of μ and σ^2 are

$$\hat{\mu} = 2\log(\hat{\mu}_1) - \frac{1}{2}\log(\hat{\mu}_2) \text{ and } \hat{\sigma}^2 = \log(\hat{\mu}_2) - 2\log(\hat{\mu}_1).$$

Can you prove that $(\hat{\mu}, \hat{\sigma}^2)$ is consistent? What about asymptotic normality? □

2 The method of maximum likelihood (ML)

The method of maximum likelihood is, by far, the most popular technique for deriving estimators and performing inference. It has an intuitive motivation and usually has fairly very good properties (at least asymptotically).

2.1 Likelihood: definition and meaning

Definition 2.1 (The likelihood function)**.** Let $f_n(\cdot;\theta)$ denote the joint pd of the sample $\boldsymbol{X} = (X_1, \ldots, X_n)$, where $\theta \in \Theta$ is the parameter of interest (θ may be vector valued; we don't bold it here for ease of notation).
Given that $\boldsymbol{X} = \boldsymbol{x} := (x_1, \ldots, x_n)$ is observed, *the function of* θ defined by

$$L_n(\theta|\boldsymbol{x}) = f_n(\boldsymbol{x};\theta),$$

is called the likelihood function of θ (given the observation $\boldsymbol{X} = \boldsymbol{x}$).

The parameter θ is listed first in L_n because, unlike f_n, L_n is viewed as a function of θ, for a given sample point $\boldsymbol{x}$ of $\boldsymbol{X}$.

If $X_1, \ldots, X_n$ are **iid** with marginal pd $f(x;\theta)$, the likelihood factorizes into

$$L_n(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i;\theta)$$

For the discrete case, $L_n(\theta|\boldsymbol{x}) = P_\theta(\boldsymbol{X} = \boldsymbol{x})$, where P_θ means that *the probability is taken under the assumption that θ is the true parameter*. So, if we compare the likelihood function at two parameter points, say θ_1 and θ_2, and find, for example, that

$$L_n(\theta_1|\boldsymbol{x}) = P_{\theta_1}(\boldsymbol{X} = \boldsymbol{x}) > L_n(\theta_2|\boldsymbol{x}) = P_{\theta_2}(\boldsymbol{X} = \boldsymbol{x}),$$

then the sample $\boldsymbol{x}$ we actually observe is *more likely* to have occurred if $\theta = \theta_1$ than if $\theta = \theta_2$. This can be interpreted by saying that, given the data $\boldsymbol{x}$, θ_1 is a more plausible value for θ than θ_2.
In general, we can think of $L_n(\theta|\boldsymbol{x})$ as a measure of how "likely" θ has produced the observed $\boldsymbol{x}$. A similar interpretation applies to the continuous case.

Example 2.1. Let X_i, $i = 1, \ldots, n$, be an iid sample from $Ber(\pi)$, where π is our parameter of interest. Suppose that we have $n = 6$ observations $x_i = 1, 1, 0, 1, 1, 0$, and $\pi \in \{0.2, 0.3, 0.7, 0.8, 0.9\} = \Theta$.
The likelihood function of the observed data is

$$L_n(\pi|\boldsymbol{x}) = P(X_1 = 1, X_2 = 1, \ldots, X_6 = 0) = \prod_{i=1}^{6} P(X_i = x_i) = \pi^4(1-\pi)^2.$$

π	$L_n(\pi)$
0.2	0.001
0.3	0.004
0.7	0.0216
0.8	0.0164
0.9	0.0066

Given the actual sample, we can say that 0.7 is the most plausible value of π among $\{0.2, 0.3, 0.7, 0.8, 0.9\}$. $\square$

Definition 2.2 (Maximum Likelihood Estimator)**.** A statistic $\hat{\theta} \equiv \hat{\theta}(\boldsymbol{X})$ is called a maximum likelihood estimator (MLE) of $\theta \in \Theta$, if (i) $\hat{\theta} \in \Theta$, and (ii) for each sample point $\boldsymbol{x}$,

$$L_n(\hat{\theta}(\boldsymbol{x})|\boldsymbol{x}) \geq L_n(\theta|\boldsymbol{x}), \ \forall \theta \in \Theta.$$

In other words, a MLE $\hat{\theta}(\boldsymbol{x})$ is a parameter point at which the likelihood $L_n(\theta|\boldsymbol{x})$, as a function of θ, attains its maximum: $L_n(\hat{\theta}(\boldsymbol{x})|\boldsymbol{x}) = \max_{\theta \in \Theta} L_n(\theta|\boldsymbol{x})$. In mathematical notation, we write

$$\hat{\theta}_n(\boldsymbol{x}) = \arg\max_{\theta \in \Theta} L_n(\theta|\boldsymbol{x}),$$

where arg max is a shortcut for "the arguments of the maxima", i.e. any point at which $\theta \mapsto L_n(\theta|\boldsymbol{x})$ is maximized over Θ.

Facts to know

- MLE *may not exist* and *may not be unique*.
- By construction, the range of MLE coincides with Θ, the range of the parameter.
- MLE is a judicious choice in that it represents the parameter point that is most likely to have generated the data. In general, this is a good point estimator in that it possesses some *optimal properties* that will be examined later.
- However, this method has the inherent drawback of having to find the maximum of the likelihood function, which is often a difficult problem. In fact, it can be challenging to find (analytically or numerically) a *global maximizer* and to ensure that it is indeed a global (and not local) maximizer, especially in the multi-parameter case; more on this later.

2.2 MLE implementation

For the sake of generality, we consider here the multi-parameter case, where the parameter of interest $\boldsymbol{\theta}$ is a vector of dimension d.

Maximum likelihood estimators are often found by maximizing the log-likelihood function

$$\ell_n(\boldsymbol{\theta}|\boldsymbol{x}) = \log(L_n(\boldsymbol{\theta}|\boldsymbol{x})) = \sum_{i=1}^{n} \log f(x_i; \boldsymbol{\theta}).$$

Since the logarithmic transformation is strictly monotonically increasing, it does not make any difference to maximize ℓ_n or L_n.

If the likelihood function is differentiable, possible candidates for the MLE are the $\boldsymbol{\theta}$'s that solve the ***likelihood equation***:

$$\partial_{\theta_k} \ell_n(\boldsymbol{\theta}|\boldsymbol{x}) = 0, \ \forall k = 1, \ldots, d.$$

This is equivalent to solve the ***Score equation*** $S_n(\boldsymbol{\theta}, \boldsymbol{x}) = \mathbf{0}$, where $S_n(\boldsymbol{\theta}, \boldsymbol{x})$ is the ***score function*** associated with $f_n(\boldsymbol{x}; \boldsymbol{\theta})$, i.e. $S_n(\boldsymbol{\theta}, \boldsymbol{x}) = (\partial_{\theta_1} \ell_n(\boldsymbol{\theta}|\boldsymbol{x}), \ldots, \partial_{\theta_d} \ell_n(\boldsymbol{\theta}|\boldsymbol{x}))^t$.

Let's first focus on the one-parameter case. When looking for a maximum using derivatives, remember that :

- ***Stationary points*** (i.e. any θ for which $\ell_n'(\theta|\boldsymbol{x}) = 0$) may be local or global minimizer of $\theta \mapsto \ell_n(\theta|\boldsymbol{x})$, local or global maximizer, or inflection points (i.e. points at which the concavity changes). Our goal is to find a ***global maximizer***.

- The ***second derivative test*** can be used to determine the nature of a stationary point (min/max/inflection). In fact, *if the second derivative at a stationary point is negative, than this point is a* ***local maximizer***. Furthermore, if $\ell_n''(\theta|\boldsymbol{x}) < 0$, $\forall \theta \in \Theta$, then ℓ_n is ***strictly concave***, and hence any stationary point that can be found will be the unique global maximizer.
- Maximum (or minimum) can occur where the derivative does not exist. $\rightarrow$ *Check all the points where the derivative dose note exists*.
- The zeros of the first derivative locate only extremum points *in the interior of the domain* of a function (here Θ). If the extremum occurs on the boundary, the first derivative may not be 0. $\rightarrow$ *Check the endpoints of* Θ.
- When all maxima candidates (if any) have been identified, the one(s) with the highest likelihood is/are the MLE.

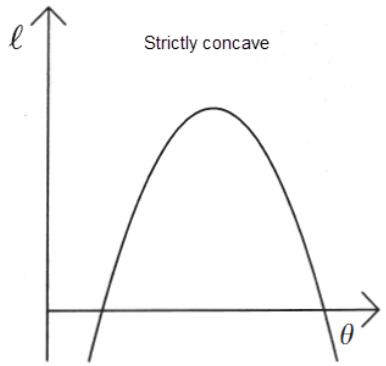
For most of the models considered here after, only one maximum exists, and it corresponds to the solution of the score equation.

Basically, the same approach can be applied for the multi-parameter case but the computational complexity increases with the number of parameters to be estimated. For example, in the case of a model with two parameters, i.e. $\boldsymbol{\theta} = (\theta_1, \theta_2)$, a stationary point, say $\boldsymbol{\theta}^*$, is a ***local maxima*** if *the Hessian matrix*

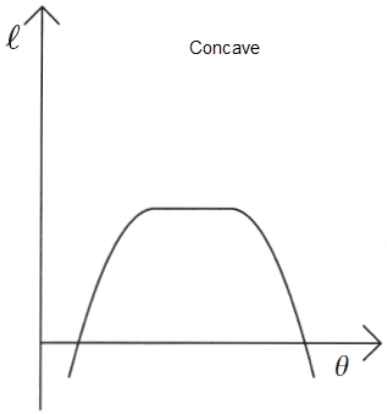
$$\boldsymbol{\nabla}^2 \ell_n(\boldsymbol{\theta}^*|\boldsymbol{x}) = \begin{pmatrix} \partial^2_{\theta_1} \ell_n(\boldsymbol{\theta}^*|\boldsymbol{x}) & \partial_{\theta_1\theta_2} \ell_n(\boldsymbol{\theta}^*|\boldsymbol{x}) \\ \partial_{\theta_1\theta_2} \ell_n(\boldsymbol{\theta}^*|\boldsymbol{x}) & \partial^2_{\theta_2} \ell_n(\boldsymbol{\theta}^*|\boldsymbol{x}) \end{pmatrix}$$

is *negative definite*. This is the case if and only if (i) $\partial^2_{\theta_1}\ell_n(\boldsymbol{\theta}^*|\boldsymbol{x}) < 0$, and (ii) $\det\left(\boldsymbol{\nabla}^2\ell_n(\boldsymbol{\theta}^*|\boldsymbol{x})\right) := \partial^2_{\theta_1}\ell_n(\boldsymbol{\theta}^*|\boldsymbol{x})\partial^2_{\theta_2}\ell_n(\boldsymbol{\theta}^*|\boldsymbol{x}) - \left(\partial_{\theta_1\theta_2}\ell_n(\boldsymbol{\theta}^*|\boldsymbol{x})\right)^2 > 0$.

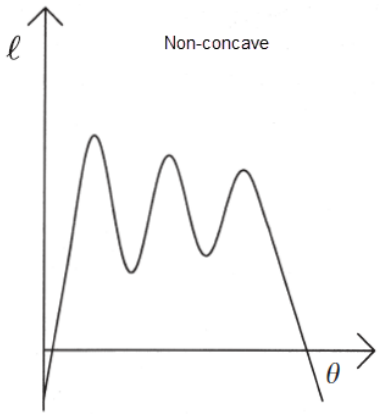
The figure below shows some log-likelihood functions and their extremum values/points in the case of one and two parameters.



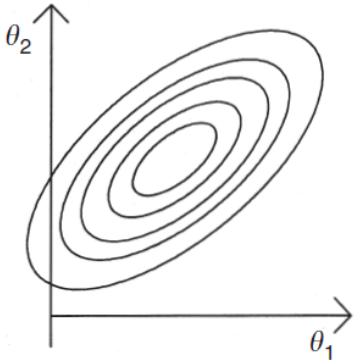
Well-behaved one-dimensional likelihood



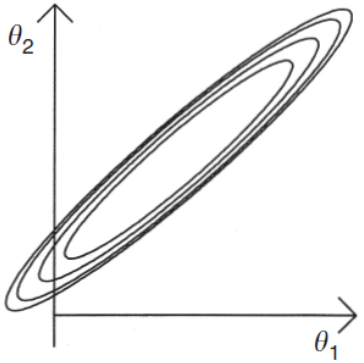
Flat one-dimensional likelihood



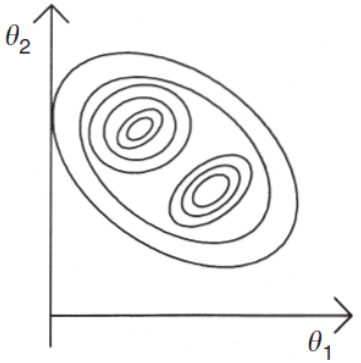
Multimodal one-dimensional likelihood



Well-behaved two-dimensional likelihood



Flat (ridged) two-dimensional likelihood

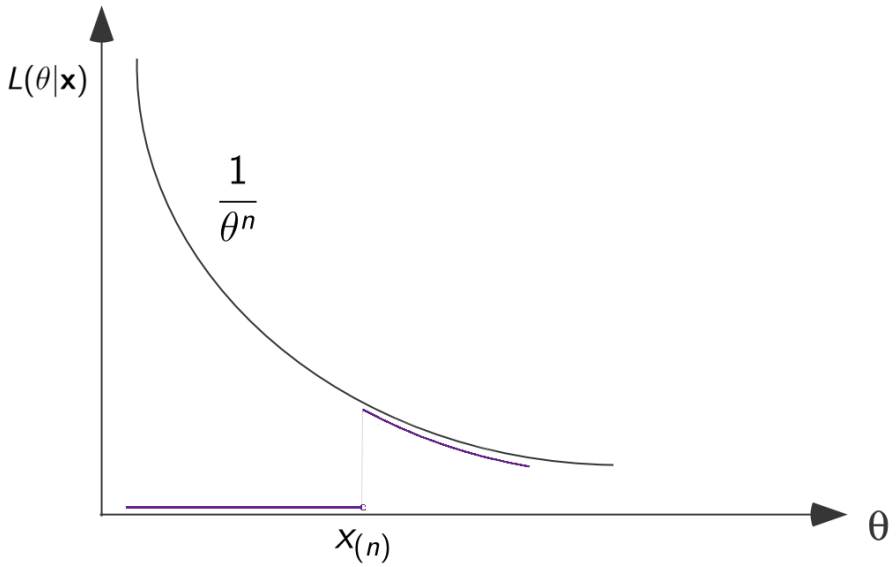


Multimodal two-dimensional likelihood

Example 2.2.

- Let X_i, $i = 1, \ldots, n$, be an iid sample from $Unif[0, \theta]$, $\theta > 0$.

$$L_n(\theta|\boldsymbol{x}) = \frac{1}{\theta^n} I(x_{(n)} \leq \theta).$$



$\implies$ $X_{(n)}$ is the MLE of θ. We have seen that $X_{(n)}$ is biased, so it is not efficient.

- Let X_i, $i = 1, \ldots, n$, be an iid sample from $Unif[\theta, \theta + 1]$, $\theta > 0$.

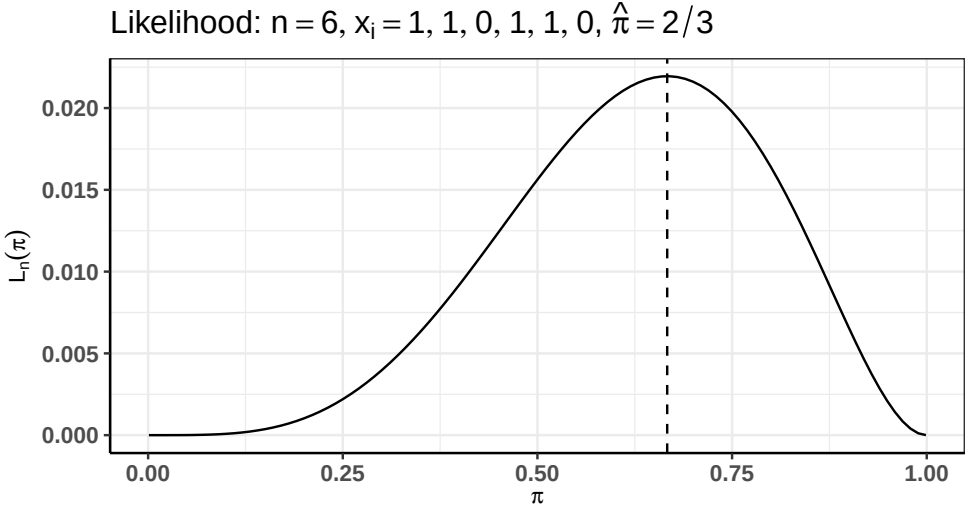
$$L_n(\theta|\boldsymbol{x}) = I(x_{(n)} - 1 \leq \theta \leq x_{(1)}).$$

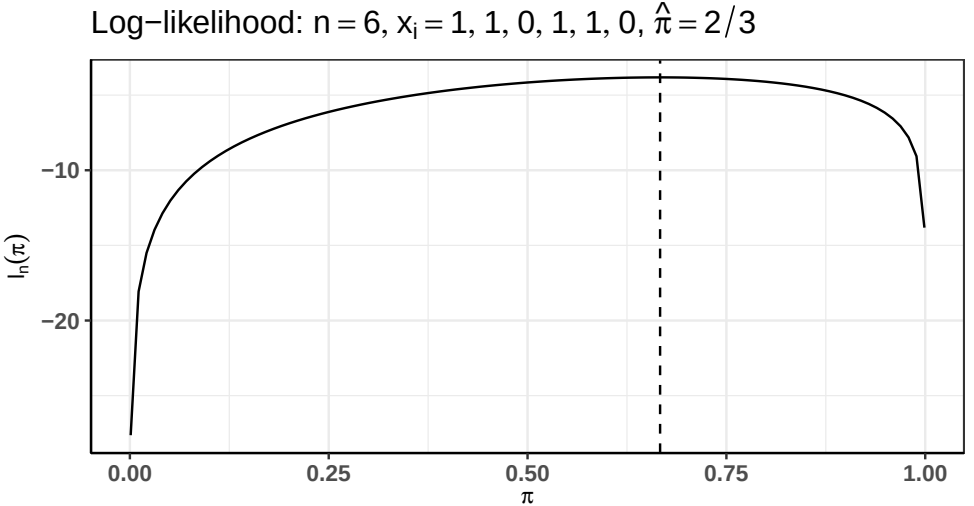
  So any $\hat{\theta}$ in $[X_{(n)} - 1, X_{(1)}]$ is a MLE estimator of θ.

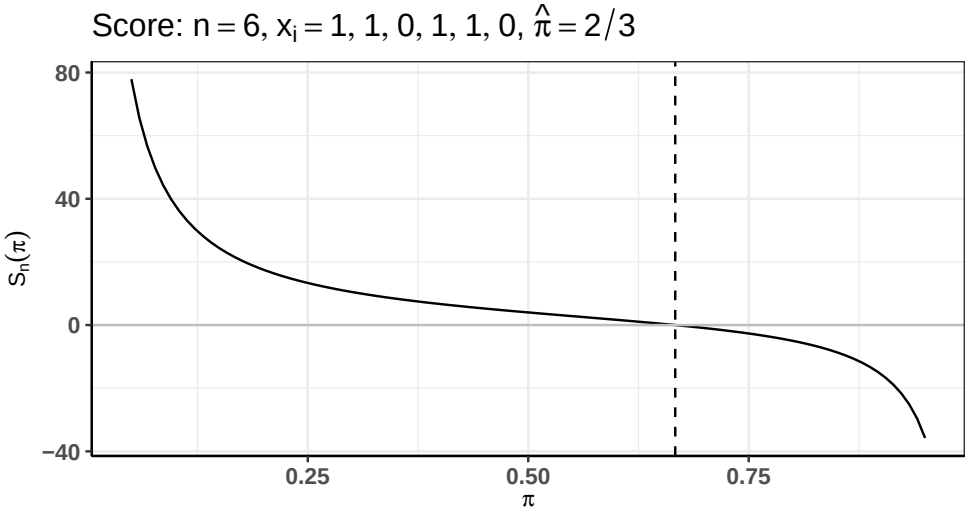
- Let X_i, $i = 1, \ldots, n$, be an iid sample from $Ber(\pi)$, $\pi \in (0, 1)$.

$$\begin{aligned}
L_n(\pi|\boldsymbol{x}) &= \pi^{\sum_{i=1}^n x_i}(1-\pi)^{n-\sum_{i=1}^n x_i} \\
\ell_n(\pi|\boldsymbol{x}) &= \left(\sum_{i=1}^n x_i\right)\log(\pi) + \left(n - \sum_{i=1}^n x_i\right)\log(1-\pi) \\
S_n(\pi, \boldsymbol{x}) &= \ell_n^{'}(\pi|\boldsymbol{x}) = \frac{\sum_i x_i}{\pi} - \frac{n - \sum_i x_i}{1-\pi} \\
S_n^{'}(\pi, \boldsymbol{x}) &= \ell_n^{''}(\pi|\boldsymbol{x}) = -\left(\frac{\sum_i x_i}{\pi^2} + \frac{n - \sum_i x_i}{(1-\pi)^2}\right).
\end{aligned}$$

  $S_n = 0 \Leftrightarrow \hat{\pi} = \frac{1}{n}\sum_{i=1}^n x_i$, and ℓ_n is strictly concave ($\ell_n^{''} < 0$, $\forall \pi$), so $\hat{\pi} = \overline{X}_n$ is the MLE of π.


Likelihood: n = 6, x_i = 1, 1, 0, 1, 1, 0, $\hat{\pi} = 2/3$
$L_n(\pi)$
0.020
0.015
0.010
0.005
0.000
0.00
0.25
0.50
0.75
1.00
π


Log−likelihood: n = 6, x_i = 1, 1, 0, 1, 1, 0, $\hat{\pi} = 2/3$
$l_n(\pi)$
−10
−20
0.00
0.25
0.50
0.75
1.00
π


Score: n = 6, x_i = 1, 1, 0, 1, 1, 0, $\hat{\pi} = 2/3$
$S_n(\pi)$
80
40
0
−40
0.25
0.50
0.75
π

- Let X_i, $i = 1, \ldots, n$, be an iid sample from $N(\mu, \sigma^2)$, $\mu \in (-\infty, \infty)$, and $\sigma^2 \in (0, \infty)$.

$$L_n(\mu, \sigma^2 | \boldsymbol{x}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$
$$\ell_n(\mu, \sigma^2 | \boldsymbol{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

  - Let's first consider the case of an unknown μ and a known σ^2. The score for μ and its derivative (with respect to μ) are

$$S_{1n} = \partial_\mu \ell_n(\mu, \sigma^2 | \boldsymbol{x}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$$
$$\partial_\mu S_{1n} = \partial_\mu^2 \ell_n(\mu, \sigma^2 | \boldsymbol{x}) = -n/\sigma^2.$$

    As a function of μ, ℓ_n is strictly concave. Hence, setting the first derivative equal to zero (i.e. $S_{1n} = 0$) and solving for μ we get $\overline{X}_n$ as the MLE of μ.

  - Now, let's consider the case of a known μ and an unknown σ^2. The score for σ^2 and its derivative (with

respect to σ^2) are

$$S_{2n} = \partial_{\sigma^2}\ell_n(\mu, \sigma^2|\boldsymbol{x}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2$$
$$\partial_{\sigma^2}S_{2n} = \partial^2_{\sigma^2}\ell_n(\mu, \sigma^2|\boldsymbol{x}) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2.$$

As a function of σ^2, ℓ_n is not strictly concave, because $\partial^2_{\sigma^2}\ell_n(\mu, \sigma^2|\boldsymbol{x})$ is not negative for all possible parameter values. Now, by setting the first derivative equal to zero (i.e. $S_{2n} = 0$) and solving for σ^2 we obtain as the *unique* solution $\tilde{\sigma}^2_n = n^{-1}\sum_{i=1}^{n}(x_i - \mu)^2$, and it is easy to see that

$$\partial^2_{\sigma^2}\ell_n(\mu, \tilde{\sigma}^2_n|\boldsymbol{x}) = -\frac{n}{2\tilde{\sigma}^4_n} < 0.$$

We conclude that $\tilde{\sigma}^2_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ is the MLE of σ^2.

– Finally, let's consider the case of unknown $\boldsymbol{\theta} = (\mu, \sigma^2)$. The score (vector) and the Hessian (matrix) are

$$S_n(\boldsymbol{\theta}, \boldsymbol{x}) = (S_{1n},\, S_{2n})$$
$$\boldsymbol{\nabla}^2_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}|\boldsymbol{x}) = \begin{pmatrix} \partial_\mu S_{1n} & \partial_{\sigma^2} S_{1n} \\ \partial_\mu S_{2n} & \partial_{\sigma^2} S_{2n} \end{pmatrix} = -\begin{pmatrix} \frac{n}{\sigma^2} & \frac{1}{\sigma^4}\sum_i (x_i - \mu) \\ \frac{1}{\sigma^4}\sum_i (x_i - \mu) & -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}\sum_i (x_i - \mu)^2 \end{pmatrix}.$$

We can see that $S_n(\hat{\boldsymbol{\theta}}, \boldsymbol{x}) = (0,\, 0) \Leftrightarrow \hat{\mu} = \overline{x}_n$, and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2$. So, $\hat{\boldsymbol{\theta}} = (\overline{x}_n,\, \hat{\sigma}^2_n)$ is the unique candidate for the MLE. Now,

$$\nabla^2_{\hat{\boldsymbol{\theta}}} \ell_n(\hat{\boldsymbol{\theta}}|\boldsymbol{x}) = -\begin{pmatrix} \frac{n}{\hat{\sigma}^2_n} & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4_n} \end{pmatrix} \prec 0 \quad \text{(i.e. negative definite).}$$

We conclude that $\hat{\boldsymbol{\theta}} = \left(\overline{X}_n,\, n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2\right)$ is the MLE of $\boldsymbol{\theta} = (\mu, \sigma^2)$.

- Let $\boldsymbol{X_i} = (X_{1i}, X_{2i}, X_{3i})$, $i = 1, \ldots, n$, be iid rve from the trinomial distribution with joint pd

$$f(\boldsymbol{x}; \boldsymbol{\pi}) = \frac{m!}{x_1! x_2! x_3!} \pi_1^{x_1} \pi_2^{x_2} \pi_3^{x_3},$$

where $m \geq 1$, $\boldsymbol{x} = (x_1, x_2)$, and $\boldsymbol{\pi} = (\pi_1, \pi_2)$ is unknown; with $x_1 = 0, \ldots, m$, $x_2 = 0, \ldots, m$, $x_3 = m - x_1 - x_2$,

$\pi_1 \in (0,1)$, $\pi_2 \in (0,1)$, and $\pi_3 = 1 - \pi_1 - \pi_2$. We have that

$$
\begin{aligned}
L_n(\pi_1, \pi_2|\boldsymbol{x}) &= \frac{(m!)^n}{\prod_{i=1}^n (x_{1i}!x_{2i}!x_{3i}!)} \pi_1^{\sum_{i=1}^n x_{1i}} \pi_2^{\sum_{i=1}^n x_{2i}} \pi_3^{\sum_{i=1}^n x_{3i}}, \\
\ell_n(\pi_1, \pi_2|\boldsymbol{x}) &= \sum_{i=1}^n x_{1i} \log(\pi_1) + \sum_{i=1}^n x_{2i} \log(\pi_2) + \sum_{i=1}^n x_{3i} \log(\pi_3) + const, \\
S_{1n} := \partial_{\pi_1} \ell_n(\pi_1, \pi_2|\boldsymbol{x}) &= \frac{\sum_i x_{1i}}{\pi_1} - \frac{\sum_i x_{3i}}{\pi_3} \text{ and} \\
S_{2n} := \partial_{\pi_2} \ell_n(\pi_1, \pi_2|\boldsymbol{x}) &= \frac{\sum_i x_{2i}}{\pi_2} - \frac{\sum_i x_{3i}}{\pi_3}.
\end{aligned}
$$

The score equation is equivalent to $\pi_3 \sum_i x_{1i} = \pi_1 \sum_i x_{3i}$ and $\pi_3 \sum_i x_{2i} = \pi_2 \sum_i x_{3i}$. Summing these two equations yields to $nm\hat{\pi}_3 = \sum_i x_{3i}$. So the unique candidate for the MLE of (π_1, π_2) is $(\hat{\pi}_1, \hat{\pi}_2)$, with $\hat{\pi}_k = \sum_i x_{ki}/nm$. The Hessian of the log-likelihood is

$$
\nabla^2_{\boldsymbol{\pi}} \ell_n(\boldsymbol{\pi}|\boldsymbol{x}) = -\begin{pmatrix} \frac{\sum_i x_{1i}}{\pi_1^2} + \frac{\sum_i x_{3i}}{\pi_3^2} & \frac{\sum_i x_{3i}}{\pi_3^2} \\ \frac{\sum_i x_{3i}}{\pi_3^2} & \frac{\sum_i x_{2i}}{\pi_2^2} + \frac{\sum_i x_{3i}}{\pi_3^2} \end{pmatrix}
$$

This matrix is negative-definite, hence ℓ_n is strictly concave. We conclude that $(\hat{\pi}_1, \hat{\pi}_2) = \left(\frac{\overline{X}_{1n}}{m}, \frac{\overline{X}_{2n}}{m}\right)$ is the MLE of (π_1, π_2). $\square$