Module 2: Statistical inference (I)

Benjamin Smith

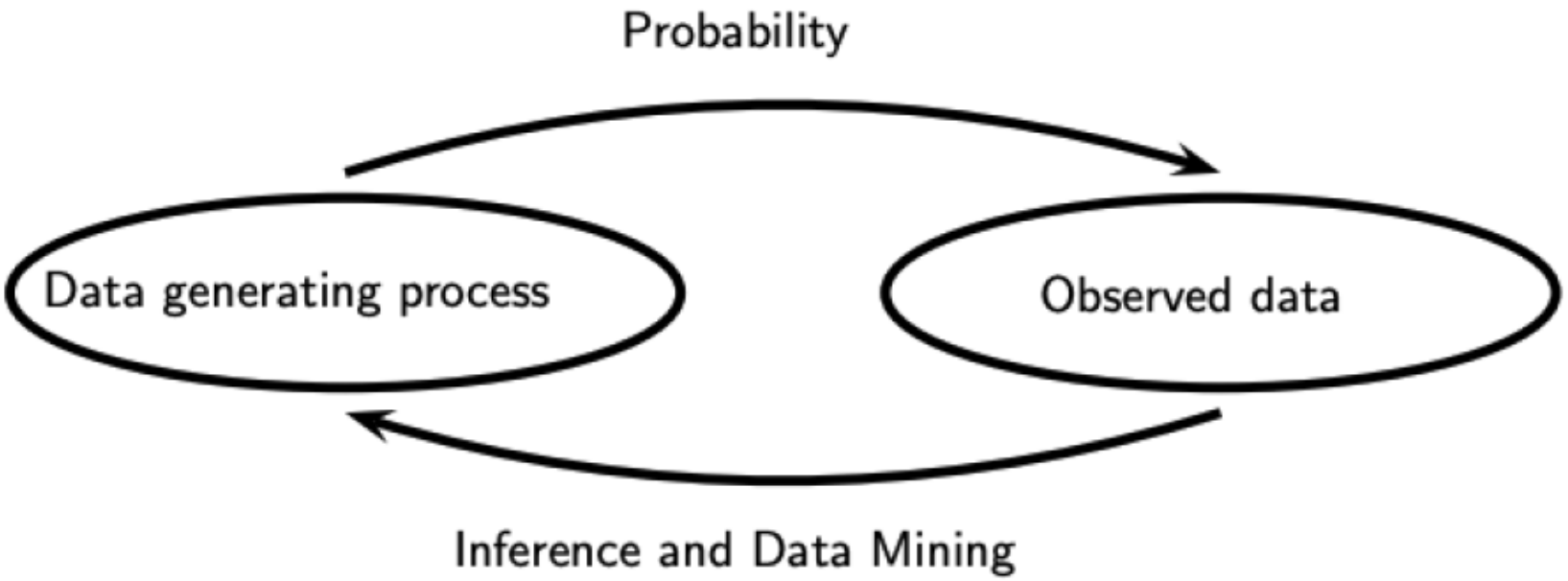
07/08/2026

Outline

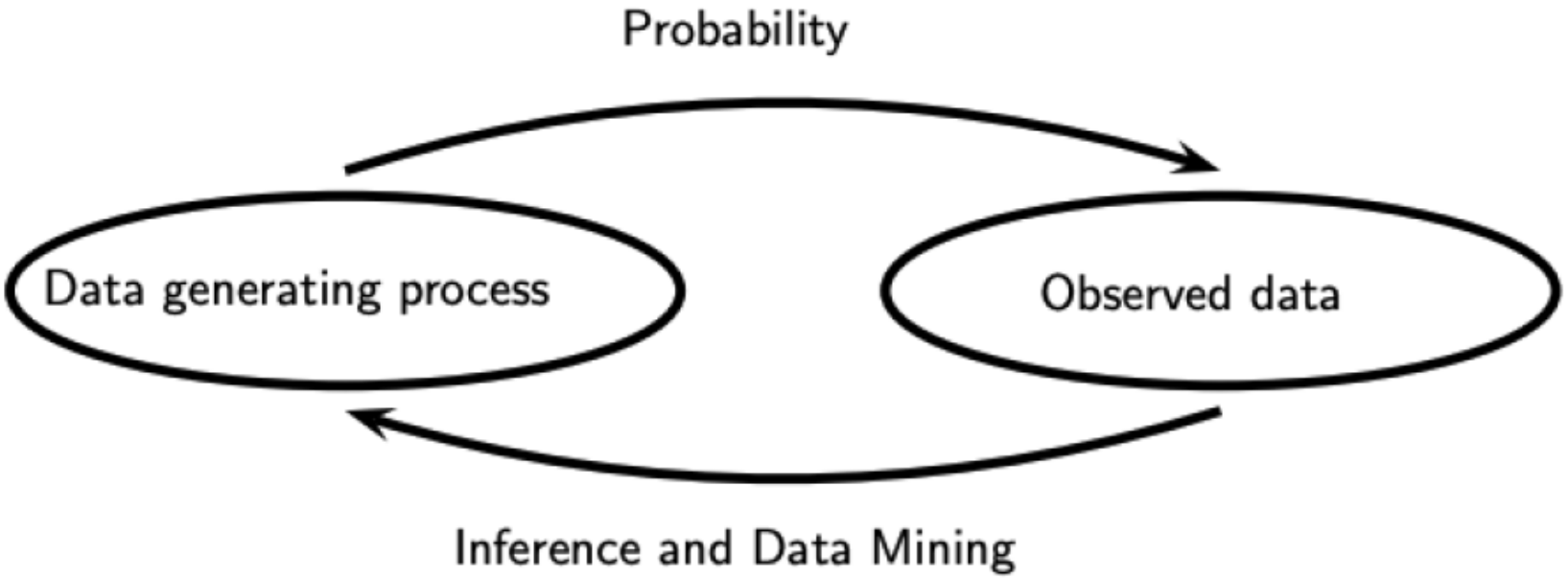
This module we will review

- Basics of parametric inference
- Methods for generating parametric estimators
- Maximum likelihood estimators
- Delta method
- Optimization method for finding MLE in R (`optim()`, Newton-Raphson)

Probability and inference



Probability and inference



- Probability: Given a data generating process, what are the properties of the outcomes?
- Statistical inference: Given the outcomes, what can we say about the process that generated the data?

Frequentist and Bayesian

- Frequentist: statistical methods with guaranteed frequency behavior
- Bayesian: statistical methods for using data to update beliefs

Point estimation

- Providing a single "best guess" of some quantity of interest
- Notations
  - Parameter θ: fixed, unknown quantity
  - Point estimator $\hat{\theta}$: depends on data, random variable

Point estimation

- Providing a single "best guess" of some quantity of interest
- Notations
  - Parameter θ: fixed, unknown quantity
  - Point estimator $\hat{\theta}$: depends on data, random variable

Definition (Point estimator)

Let $X_1, \ldots, X_n$ be n IID data points from some distribution F. A point estimator $\hat{\theta}_n$ of a parameter θ is some function of $X_1, \ldots, X_n$:

$$\widehat{\theta}_n = g\left(X_1, \ldots, X_n\right)$$

- What is a good point estimate?

MSE

- Definition:

$$\mathrm{MSE} = \mathbb{E}_\theta \left(\widehat{\theta}_n - \theta \right)^2$$

- No uniformly best estimator in terms of MSE
- It is NOT possible to have an estimator that is uniformly the best.

Bias and Variance

- Bias

$$\text{bias}\left(\widehat{\theta}_n\right) = \mathbb{E}_\theta\left(\widehat{\theta}_n\right) - \theta$$

- Variance

$$\text{Var}\left(\widehat{\theta}_n\right) = \mathbb{E}_\theta\left(\widehat{\theta}_n - \mathbb{E}\theta\right)^2$$

- Theorem

$$MSE = bias^2 + Var$$

Unbiasedness

- Definition

$$\text{bias}\left(\widehat{\theta}_n\right) = \mathbb{E}_\theta\left(\widehat{\theta}_n\right) - \theta = 0$$

- Unbiasedness is a small sample (finite sample) property
- An unbiased estimator may not exist
- An unbiased estimator is not necessarily a good estimator

Consistency

- Definition

$$\widehat{\theta}_n \xrightarrow{\mathrm{P}} \theta$$

- It is possible to be unbiased but not consistent.
- It is possible to be consistent but not unbiased.

Asypototic unbiasedness

- Definition

$$\text{bias}\left(\widehat{\theta}_n\right) = \mathbb{E}_\theta\left(\widehat{\theta}_n\right) - \theta \to 0, \text{ as } n \to \infty$$

- It is possible to be asypototically unbiased but not consistent.
- It is possible to be consistent but NOT asymptotically unbiased.
- Sufficient conditions: $MSE \to 0$.

Parameter Estimation

Definition (Parametric models)

$$\mathfrak{F} = \{f(x;\theta) : \theta \in \Theta\}$$

where the $\Theta \subset \mathbb{R}^k$ is the parameter space and $\theta = (\theta_1, \ldots, \theta_k)$ is the parameter.

Goal of parametric inference

- estimate the parametric θ (assume we known the form of the density).

Parameter of interest

Often, we are interested in estimating some function $T(\theta)$.

For example, if $X \sim N(\mu, \sigma^2)$, then

- Parameters: $\theta = (\mu, \sigma)$
- Parameter space: $\Theta = \{(\mu, \sigma) : \mu \in \mathbb{R}, \sigma > 0\}$

If the goal is to estimate the μ then

- Parameter of interest: $T(\theta) = \mu$
- Nuisance parameter: σ

Maximum likelihood

- Parametric model: $f(x;\theta)$, $X_1, \ldots, X_n$ iid
- Likelihood function

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f\left(X_i;\theta\right)$$

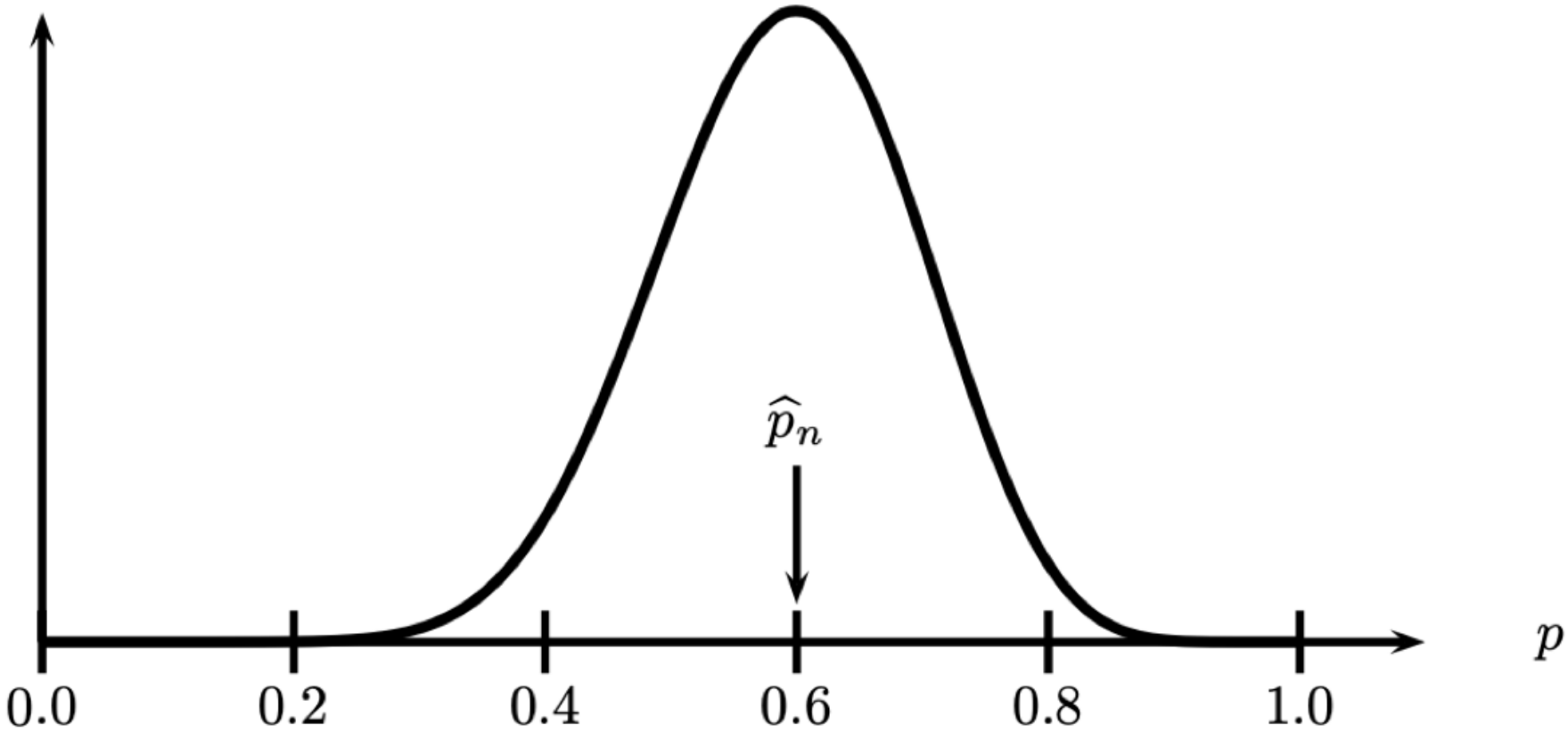
- The log-likelihood function

$$\ell_n(\theta) = \log \mathcal{L}_n(\theta) = \sum_{i=1}^{n} \log f\left(X_i;\theta\right)$$

- The maximum likelihood estimator (MLE)

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \mathcal{L}(\theta)$$

An example of MLE



Likelihood function for Bernoulli with $n = 20$ and $\sum_{i=1}^{n} X_i = 12$. The MLE is $\widehat{p}_n = 12/20 = 0.6$.

Steps to find the MLE

1. Write out the likelihood

$$\mathcal{L}(\theta) = f\left(X_1, \ldots, X_n; \theta\right)$$

2. Simplify the log likelihood

$$\ell(\theta) = \log \mathcal{L}(\theta)$$

3. Take the derivative of $\ell(\theta)$ with respect to the parameter of interest, θ
   Set $= 0$
4. Solve for θ (get $\hat{\theta}_{MLE}$)
5. Check that $\hat{\theta}_{MLE}$ is a maximum $\left(\frac{\partial^2}{\partial \theta^2}\ell(\theta) < 0\right)$

Exercise

Suppose we have an iid sample $\{X_1, \ldots, X_n\}$ with $X_i \sim \text{Bernoulli}(p)$. Find the MLE for p.

Exercise

Suppose we have an iid sample $\{X_1, \ldots, X_n\}$ with $X_i \sim \text{Bernoulli}(p)$. Find the MLE for p.

1. The likelihood

$$\mathcal{L}_n(p) = \prod_{i=1}^{n} f(X_i; p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$$

where $S = \sum_i X_i$

Exercise

Suppose we have an iid sample $\{X_1, \ldots, X_n\}$ with $X_i \sim \text{Bernoulli}(p)$. Find the MLE for p.

1. The likelihood

$$\mathcal{L}_n(p) = \prod_{i=1}^{n} f(X_i; p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$$

where $S = \sum_i X_i$

2. Log-likelihood

$$\ell_n(p) = S \log p + (n - S) \log(1 - p)$$

Exercise

Suppose we have an iid sample $\{X_1, \ldots, X_n\}$ with $X_i \sim \text{Bernoulli}(p)$. Find the MLE for p.

1. The likelihood

$$\mathcal{L}_n(p) = \prod_{i=1}^{n} f(X_i; p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$$

where $S = \sum_i X_i$

2. Log-likelihood

$$\ell_n(p) = S \log p + (n - S) \log(1 - p)$$

3. MLE

$$\ell_n'(p) = 0$$

The MLE is $\widehat{p}_n = S/n$.

Asymptotics of MLE

Under mild regularity conditions, MLEs are

- *Consistent* $\rightarrow$ converge to the true value in probability as $n \rightarrow \infty$, i.e.

$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| \leq \epsilon) = 1 \quad \forall \epsilon > 0$$

- *Asymptotically normal* $\rightarrow \sqrt{n}(\hat{\theta} - \theta) \sim N\left(0, \sigma^2\right)$ for large n
- *Asymptotically efficient*
- *equivariant* $\rightarrow$ if $\hat{\theta}$ is the MLE for θ then $g(\hat{\theta})$ is the MLE for $g(\theta)$

Asymptotic Efficiency

Cramér–Rao bound

The variance of any *unbiased* estimator $\hat{\theta}$ of θ is bounded by the reciprocal of the Fisher information $I(\theta)$:

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)},$$

where $I(\theta) = n\mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right]$.

Both MoM estimators are asymptotically unbiased, but MLE estimators achieves the CR lower bound.

MLE in R

Sometimes, there is no closed-form solution, so we need to use optimization methods to find the maximum of the log-likelihood.

- `optim()` find values of some parameters that **minimizes** some function.
- Newton-Raphson
- EM-algorithm (Not discussed here)

Example using optim()

```r
set.seed(42) # For reproducibility
sample_data <- rbinom(1000, size = 1, prob = 0.3) # Assuming success probability of 0.3

# Log-likelihood function for Bernoulli distribution
log_likelihood_bernoulli <- function(p, data) {
  n <- length(data)
  log_likelihood <- sum(data * log(p) + (1 - data) * log(1 - p))
  return(-log_likelihood) # Negative to be used with optimization functions (minimization)
}

# Initial parameter value for optimization (probability of success)
initial_param <- 0.8

# Find MLE using optim
result <- optim(
  par = initial_param, fn = log_likelihood_bernoulli,
  data = sample_data, method = "Brent", lower = 0, upper = 1
) # If bounds aren't provided, use method="BFGS"


# MLE estimate of p
mle_p <- result$par

# Print the result
cat("MLE of p:", mle_p, "\n")
```

```
## MLE of p: 0.293
```

Newton-Raphson

Derivative of the log-likelihood around θ^j :

$$0 = \ell'(\widehat{\theta}) \approx \ell'\left(\theta^j\right) + \left(\widehat{\theta} - \theta^j\right)\ell''\left(\theta^j\right)$$

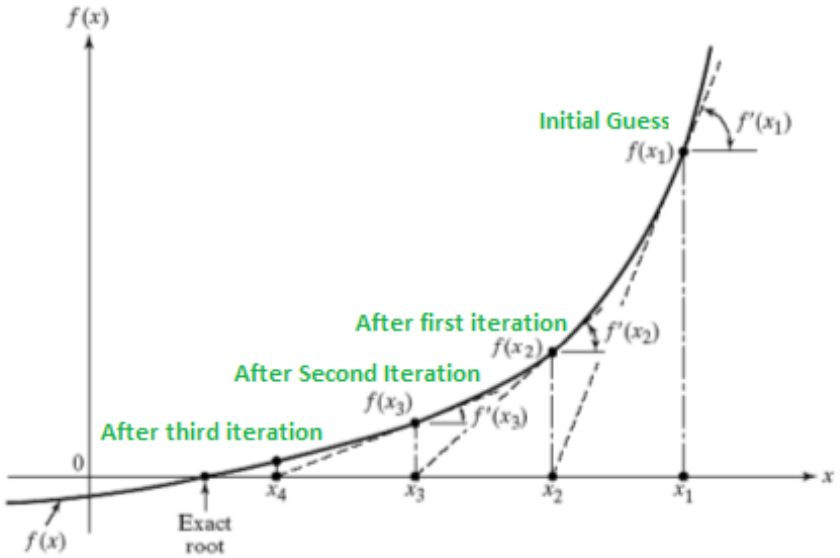
Solving for $\widehat{\theta}$ gives

$$\widehat{\theta} \approx \theta^j - \frac{\ell'\left(\theta^j\right)}{\ell''\left(\theta^j\right)}.$$

This suggests the following iterative scheme:

$$\widehat{\theta}^{j+1} = \theta^j - \frac{\ell'\left(\theta^j\right)}{\ell''\left(\theta^j\right)}$$

Illustration

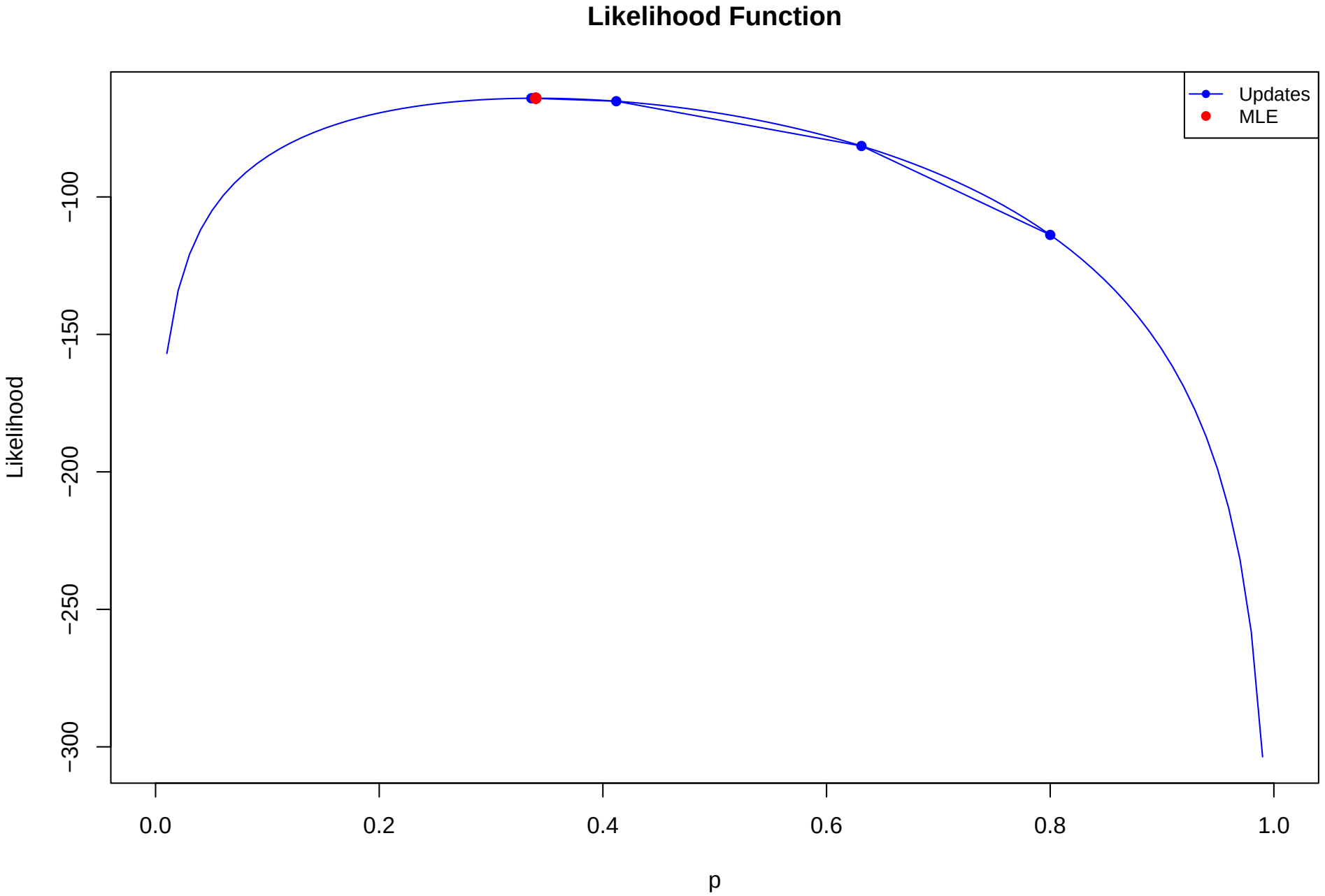


NR algorithm in R

```r
# First derivative of the log-likelihood function
log_likelihood_bernoulli_prime <- function(p, data) {
  n <- length(data)
  d_log_likelihood <- sum(data / p - (1 - data) / (1 - p))
  return(-d_log_likelihood) # Negative to be used with optimization functions (minimization)
}
# Second derivative of the log-likelihood function
log_likelihood_bernoulli_double_prime <- function(p, data) {
  n <- length(data)
  dd_log_likelihood <- sum(-data / p^2 - (1 - data) / (1 - p)^2)
  return(-dd_log_likelihood) # Negative to be used with optimization functions (minimization)
}
# Initial parameter value for optimization (probability of success)
initial_param <- 0.8
# Newton-Raphson algorithm for optimization
tolerance <- 1e-8
max_iterations <- 1000
p <- initial_param
for (i in 1:max_iterations) {
  p_new <- p - log_likelihood_bernoulli_prime(p, sample_data) /
    log_likelihood_bernoulli_double_prime(p, sample_data)
  if (abs(p_new - p) < tolerance) {
    break
  }
  p <- p_new
}
# Print the result
cat("MLE of p:", p, "\n")
```

```
## MLE of p: 0.293
```

Solution Trajectory



Resources

This tutorial is based on

- Harvard Biostatistics Summer Pre Course [link]
- "All of Statistics" by Larry A. Wasserman [link]
- "Beyond Multiple Linear Regression" by Paul Roback and Julie Legler [link]

Exercises

Available on course website.