

# **MATH60604A**

## **Statistical modelling**

### **§3.a - Maximum likelihood estimation**

Léo Belzile

HEC Montréal  
Department of Decision Sciences

# Introduction to likelihood

- Suppose we want to estimate the probability that an event occurs, which we assume is constant.
- For example, whether a customer buys a product or not, whether a study participant completes a task or not, etc.
- We have a sample size of  $n$  with  $X_i$  assumed to come from a Bernoulli distribution with probability  $p$ , meaning

$$P(X_i = 1) = p, \quad P(X_i = 0) = 1 - p.$$

- By convention, “1” denotes a success and “0” a failure.

# Joint probability of outcomes in Bernoulli example

A compact way of writing the mass function is

$$P(X_i = x_i \mid p) = p^{x_i}(1 - p)^{1-x_i}, \quad x_i \in \{0, 1\}.$$

Since the observations are independent, the joint probability of a given result is the product of the probabilities for each observation,

$$\begin{aligned} P(X_1 = x_1, \dots, X_n = x_n \mid p) &= \prod_{i=1}^n P(X_i = x_i \mid p) \\ &= \prod_{i=1}^n p^{x_i}(1 - p)^{1-x_i}. \end{aligned}$$

# Likelihood

- The **likelihood**  $L(\theta)$  is a function of the **parameters** of the distribution, say  $\theta$ .
  - The likelihood gives the probability of observing a sample under a postulated distribution whose parameters are  $\theta$ .
  - The likelihood treats the observations as fixed.
- The **maximum likelihood** estimator  $\hat{\theta}$  is the value of  $\theta$  that maximizes the likelihood.
  - the value that makes the observed sample the most **likely** or **plausible**.
  - scientific thinking: whatever we observe, we have expected to observe.

# Likelihood of the Bernoulli model

The likelihood for a random sample is

$$\begin{aligned} L(p; \mathbf{X}) &\equiv L(p; \mathbf{X}) = \prod_{i=1}^n p^{X_i} (1-p)^{(1-X_i)} \\ &= p^{\sum_{i=1}^n X_i} (1-p)^{n - \sum_{i=1}^n X_i}. \end{aligned}$$

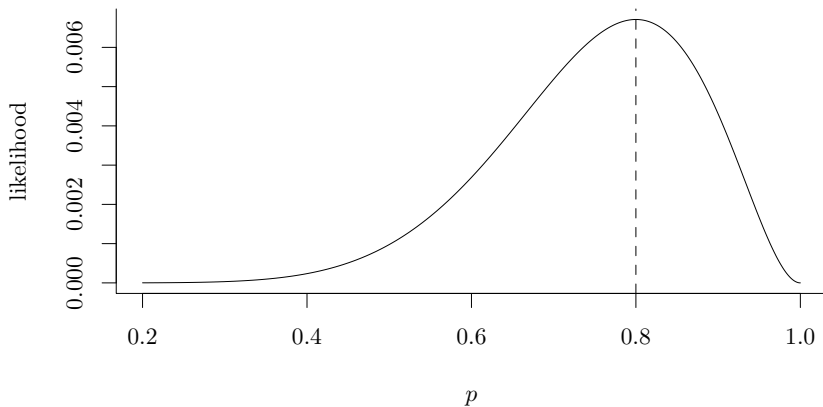
This likelihood is (up to normalizing constant) the probability of a binomial sample of size  $n$  with probability of success  $p$ .

- It is easy to see that the likelihood only depends on the number of ones observed, regardless of the ordering.
- The maximum likelihood estimator is the observed proportion of ones in the sample.

Suppose that we have  $n = 10$  observations, eight of which are successes.

- The likelihood is  $L(p) = p^8(1-p)^2$  and the maximum likelihood estimate  $\hat{p}$  is 0.8.

## Plot of the likelihood function $L(p)$



# Log-likelihood for Bernoulli sample

- In our example, the log-likelihood function is

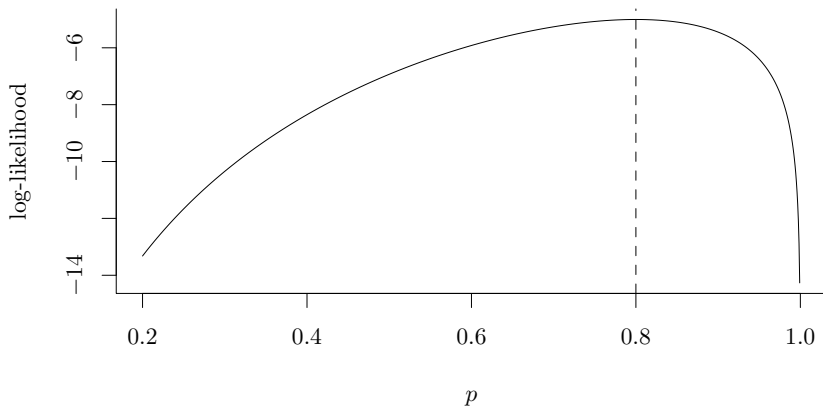
$$\ell(p) = \sum_{i=1}^n \log \left\{ p^{x_i} (1-p)^{1-x_i} \right\}$$

- Using the property  $\log(a^b) = b \log(a)$ , rewrite

$$\ell(p) = \log(p) \sum_{i=1}^n x_i + \log(1-p) \left( n - \sum_{i=1}^n x_i \right).$$

- In our numerical example, with eight ones and two zeros, the log-likelihood is  $\ell(p) = 8 \log(p) + 2 \log(1-p)$ .

## Plot of the log-likelihood function $\ell(p)$





# Maximum likelihood estimator

Differentiating  $\ell(p)$  with respect to  $p$ ,

$$\frac{d}{dp}\ell(p) = \frac{1}{p} \sum_{i=1}^n x_i - \frac{1}{(1-p)} \left( n - \sum_{i=1}^n x_i \right).$$

Solving the score equation  $U(p) = d\ell(p)/dp = 0$ , we find

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}.$$

The second derivative,

$$\frac{d^2\ell(p)}{dp^2} = -\frac{1}{p^2} \sum_{i=1}^n x_i - \frac{1}{(1-p)^2} \left( n - \sum_{i=1}^n x_i \right),$$

is negative, so  $L(p)$  thus achieves a maximum at  $\hat{p}$  and the maximum likelihood estimator of  $p$  is the sample **proportion of ones**.

# Estimating the mean and variance of a normal sample

- Suppose we have an independent normal sample of size  $n$  with mean  $\mu$  and variance  $\sigma^2$ , where

$$X_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2).$$

- The vector of parameters is  $\theta = (\mu, \sigma^2)$ .
- Recall that the density of the normal distribution is

$$f(x \mid \theta) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}, \quad x \in \mathbb{R}.$$

## Example : estimating the mean and variance of a normally-distributed population

- For a sample  $\mathbf{X} = \mathbf{x}$ , the likelihood is

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

- The log-likelihood is

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

## Analytic expression for MLE of normal sample

This is another example for which we're able to find the maximum likelihood estimator analytically. The **score equations** are

$$\frac{\partial}{\partial \mu} \ell(\boldsymbol{\theta}) = 0, \quad \frac{\partial}{\partial \sigma^2} \ell(\boldsymbol{\theta}) = 0.$$

One can show that

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2,$$

are the maximum likelihood estimators for the two parameters.

# Estimating the mean and variance of a normal distribution

- Intuitively, we would expect the estimator of the theoretical mean  $\mu$  to just be the sample mean.
- However, the estimator of  $\sigma^2$  is not the sample variance estimator,

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- In one case we divide by  $n$  (maximum likelihood estimator); in the other we divide by  $(n-1)$ .
- The two estimators are **consistent**, i.e., the estimate will get arbitrarily close to the true value  $\sigma^2$  as  $n \rightarrow \infty$ .

# Unbiasedness

- An estimator of  $\theta_i$  is **unbiased** if its expectation is equal to  $\theta_i$ .
  - On average, the estimate is centered at the true value, regardless of the sample size  $n$ .
- The maximum likelihood estimator for the mean of a normal distribution is **unbiased**, meaning that

$$E(\hat{\mu}) = \mu$$

- One can show that  $E(S^2) = \sigma^2$  and so the sample variance estimator is unbiased.
- Since  $\hat{\sigma}^2 = (n-1)/n S^2$ , it follows that the maximum likelihood estimator of  $\sigma^2$  is **biased**.

# Ordinary linear regression

Assuming normality of the errors, the least square estimators of  $\beta$  coincide with the maximum likelihood estimator of  $\beta$ .

- Recall the linear regression model,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_i, \quad (i = 1, \dots, n),$$

where the errors  $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ .

- The linear model has  $p + 2$  parameters:  $\beta_0, \beta_1, \dots, \beta_p$  and  $\sigma^2$ .
- The log-likelihood is

$$\begin{aligned} \ell(\boldsymbol{\theta}) = & -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) \\ & - \frac{1}{2\sigma^2} \sum_{i=1}^n \left( Y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij} \right)^2. \end{aligned}$$

# Least squares and maximum likelihood estimator

- Maximizing the log-likelihood with respect to  $\beta_0, \dots, \beta_p$  is equivalent to minimizing the sum of squared errors,

$$\sum_{i=1}^n \left( Y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij} \right)^2 .$$

- This objective function is the same as that of least squares.
- The least-square estimator  $\hat{\beta}$  of  $\beta$  is the maximum likelihood estimator.



# Maximum likelihood estimator of the variance in linear regression

- The maximum likelihood estimator (MLE) of the variance  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \left( Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_p X_{ip} \right)^2$$

- The usual estimator of  $\sigma^2$  is

$$S^2 = \frac{SS_e}{n - p - 1},$$

where  $p + 1$  is the number of  $\beta$ 's and  $SS_e$ , the sum of squared residuals, is

$$SS_e = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n \left( Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_p X_{ip} \right)^2.$$

- $S^2$  is unbiased for  $\sigma^2$ , unlike  $\hat{\sigma}^2$ . Both estimators are consistent.