MATH60604A
Statistical modelling
§3 - Maximum likelihood estimation

Léo Belzile

HEC Montréal
Department of Decision Sciences

Likelihood

- The likelihood $L(\theta)$ is a function of the **parameters** of the distribution, say θ.
  - The likelihood gives the probability of observing a sample under a postulated distribution whose parameters are θ.
  - The likelihood treats the observations as fixed.
- The maximum likelihood estimator $\widehat{\theta}$ is the value of θ 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.

Bernoulli trials

- 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

$$\mathsf{P}\left(X_i = 1\right) = p, \qquad \mathsf{P}\left(X_i = 0\right) = 1 - p.$$

- 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

$$\mathsf{P}\left(X_i = x_i \mid p\right) = p^{x_i}(1-p)^{1-x_i}, \qquad x_i \in \{0,1\}.$$

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

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

Likelihood of the Bernoulli model

The likelihood for the random sample is

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

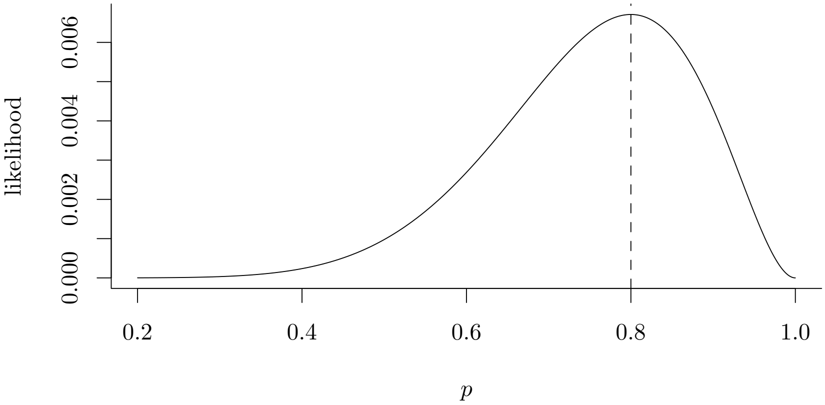
This likelihood is (up to normalizing constant) the same as that of a binomial sample of size n with probability of success p.

- the likelihood only depends on the number of successes, regardless of the ordering.

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

- The likelihood is $L(p) = p^8(1-p)^2$.

Plot of the likelihood function $L(p)$



Log-likelihood for Bernoulli sample

- The log-likelihood function is

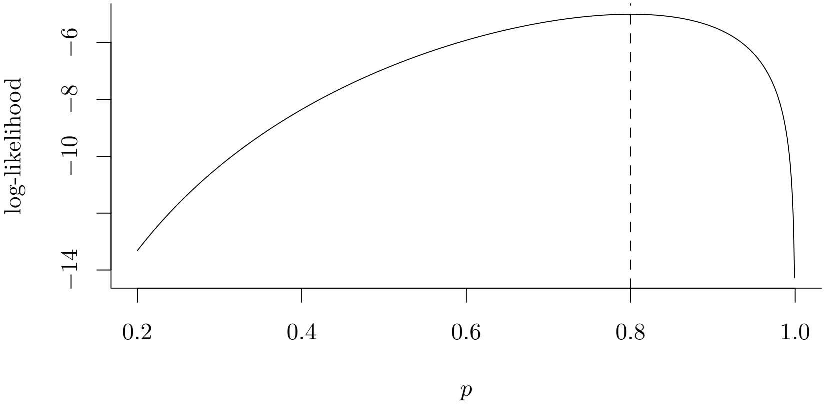
$$\ell(p) = \sum_{i=1}^{n} \ln\left\{p^{x_i}(1-p)^{1-xi}\right\}$$

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

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

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

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



Maximum likelihood estimator

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

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

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

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^n x_i = \overline{x}.$$

The second derivative,

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

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

Information

The observed information $j(p) = -\mathrm{d}^2\ell(p)/\mathrm{d}p^2$ and

$$j(\widehat{p}) = \frac{n}{\overline{x}} + \frac{n}{(1-\overline{x})} = \frac{n}{\overline{x}(1-\overline{x})}$$

so, the estimated variance of $\widehat{p}$ is $j^{-1}(\widehat{p}) = 0.016$ and the standard error 0.1265.
The Fisher information is

$$i(\theta) = \frac{n}{p(1-p)}.$$

- For independent and identically distributed data, the total information in the sample is n times that of an individual observation (information accumulates linearly).

Testing procedure and confidence interval

Suppose we are interested in the two-sided hypothesis

$$\mathscr{H}_0 : p_0 = 0.5 \qquad \text{versus} \qquad \mathscr{H}_a : p_0 \neq 0.5.$$

The three likelihood-based tests for this hypothesis are:

- the Wald test

$$W(p_0) = \frac{(\widehat{p} - p_0)^2}{\mathrm{Var}\,(\widehat{p})} = \frac{(\widehat{p} - p_0)^2}{\widehat{p}(1-\widehat{p})/n}$$

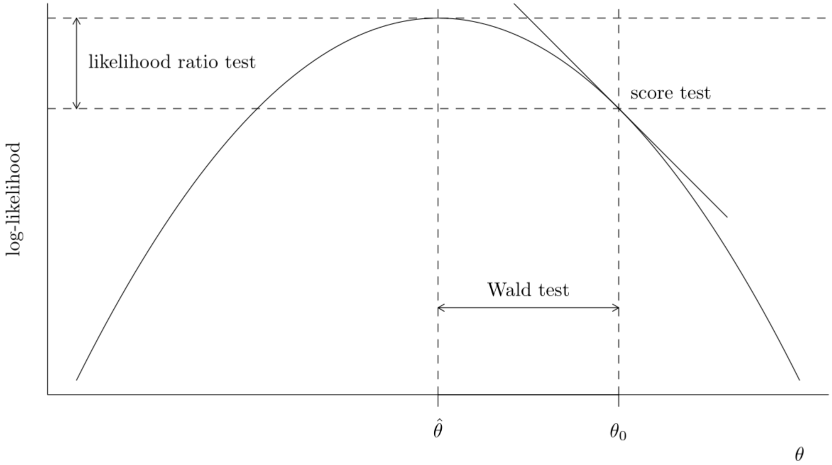
- the score test

$$S(p_0) = \frac{U^2(p_0)}{i(p_0)} = \frac{(\widehat{p} - p_0)^2}{p_0(1-p_0)/n}$$

- the likelihood ratio test

$$\begin{aligned} R(p_0) &= 2\{\ell(\widehat{p}) - \ell(p_0)\} \\ &= 2\left\{ y \ln\left(\frac{\widehat{p}}{p_0}\right) + (n-y)\ln\left(\frac{1-\widehat{p}}{1-p_0}\right)\right\} \end{aligned}$$

Illustration of likelihood-based tests



Numerical results and confidence intervals

- With 8 successes out of 10 trials, the statistics equal $W = 5.62$, $S = 3.6$, $R = 3.855$;
- we compare these values with the 0.95 quantile of the χ^2_1 distribution, 3.84.

- In small sample size or when the sampling distribution is strongly asymmetric, the Wald test is unreliable.
- Inverting the Wald statistic gives a 95% confidence interval

$$\widehat{p} \pm \mathfrak{z}_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

- The 95% Wald-based confidence interval is $0.8 \pm 1.96 \cdot 0.1265 = [0.55, 1.048]$!
- Compare with
  - the likelihood ratio test confidence interval is $[0.5005, 0.964]$.
  - the score test confidence interval is $[0.49, 0.943]$.

Solve $\{p : S(p) \leq 3.84\}$ and $\{p : R(p) \leq 3.84\}$ via root finding.

Estimating the mean and variance of a normal sample

- Suppose we have an independent normal sample of size n with mean μ and variance σ^2, where

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

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

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

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

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

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

- The log-likelihood is

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

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

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

One can show that

$$\widehat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

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 μ to just be the sample mean.
- However, the estimator of σ^2 is not the sample variance estimator,

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

- 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 σ^2 as $n \to \infty$.

Unbiasedness

- An estimator of θ_i is **unbiased** if its expectation is equal to θ_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

$$\mathsf{E}\left(\widehat{\mu}\right) = \mu$$

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

Ordinary linear regression

Assuming normality of the errors, the least square estimators of β coincide with the maximum likelihood estimator of β.

- Recall the linear regression model,

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

  where the errors $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.
- The linear model has $p + 2$ parameters: $\beta_0, \beta_1, \ldots, \beta_p$ and σ^2.
- The log-likelihood is

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

Least squares and maximum likelihood estimator

- Maximizing the log-likelihood with respect to $\beta_0, \ldots, \beta_p$ 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 $\widehat{\beta}$ of β is the maximum likelihood estimator.

Maximum likelihood estimator of the variance in linear regression

- The maximum likelihood estimator (MLE) of the variance σ^2 is

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

- The usual estimator of σ^2 is

$$S^2 = \frac{\mathrm{SS}_e}{n-p-1},$$

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

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

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