MATH60604A
Statistical modelling
§3 - Likelihood-based inference

HEC Montréal
Department of Decision Sciences

Likelihood

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

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

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

- 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

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

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

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

Likelihood of the Bernoulli model

The likelihood for the random sample is

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

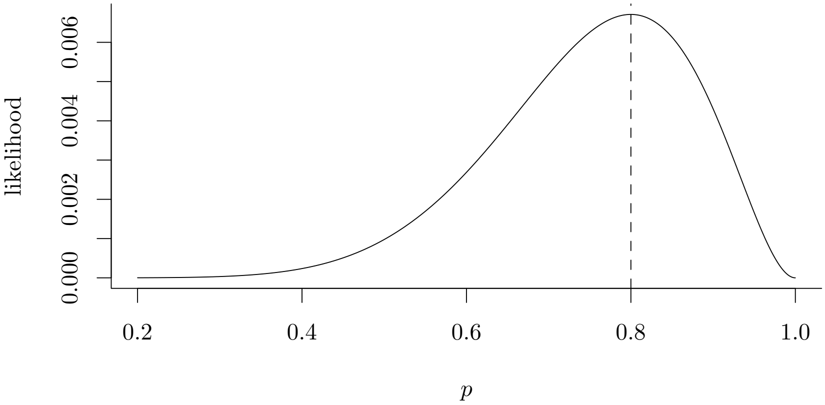
This likelihood is (up to normalizing constant) the same as that of a binomial sample of size n with probability of success p.

- 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

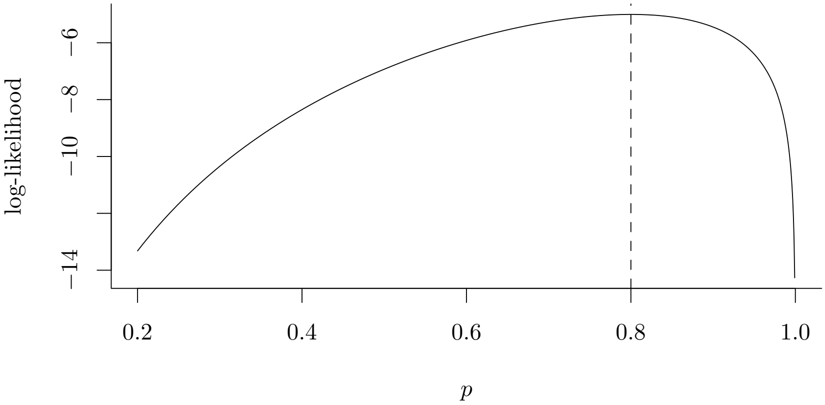
$$\ell(p) = \sum_{i=1}^{n} \ln \left\{ p^{x_i} (1-p)^{1-xi} \right\}$$

- 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

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

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

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

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

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

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

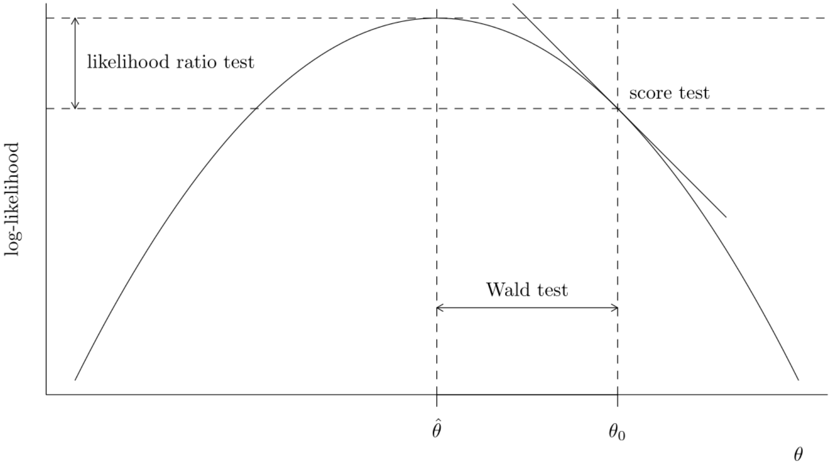
- the score test

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

- 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 95% confidence intervals based on
  - the likelihood ratio statistic, $[0.5005, 0.964]$.
  - the score statistic, $[0.49, 0.943]$.

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