Max likelihood estimators

MLE a method of estimating the parameters of a statistical model based on optimization of a likelihood function.

Suppose that random variables $X_1, \ldots, X_n$ form a random sample from a distribution $f(x|\theta)$ where θ is the parameter of interest (e.g. $\theta = \mu$).

Ex) Suppose we have a random sample

$X_1, X_2, \ldots, X_n$ of students:

$X_i = 0$ if student i does not own a dog

$X_i = 1$ if student i owns a dog.

Assuming X_i are independent Bernoulli random variables with parameter p (unknown), find the max likelihood estimator of p.

(proportion of students who own dogs).

For each student i, the probability mass function:

$$f(x_i, p) = p^{x_i}(1-p)^{1-x_i} = P(X_i = x_i)$$

for $x_i = 0$ or $x_i = 1$ and $0 \leq p \leq 1$.

The likelihood function $L(p)$ is defined as:

$$L(p) = \prod_{i=1}^{n} f(x_i, p) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

$$= p^{x_1}(1-p)^{1-x_1} \cdot p^{x_2}(1-p)^{1-x_2} \cdot \ldots \cdot p^{x_n}(1-p)^{1-x_n}$$

$$= p^{\sum x_i}(1-p)^{n-\sum x_i}$$

We must find p that maximizes the likelihood $L(p)$.

The MLE estimate of p is the value of p under which your observed data most likely fall.

It is often easier to maximize the log likelihood. The p that maximizes $\log L(p)$ is also the p which maximizes $L(p)$.

$$\log L(p) = \left(\sum x_i\right) \log(p) + \left(n - \sum x_i\right) \log(1-p)$$

Various ways to maximize this.

Using calculus:

$$\frac{\partial \log L(p)}{\partial p} = \frac{\sum x_i}{p} - \frac{\left(n - \sum x_i\right)}{1-p} = 0$$

$$\Rightarrow \left(\sum x_i\right)(1-p) - \left(n - \sum x_i\right)p = 0$$

$$\Rightarrow \sum x_i - p\sum x_i - np + p\sum x_i = 0$$

$$\Rightarrow \hat{p} = \frac{\sum x_i}{n} \quad \text{(as expected)}$$

Using R

```
p = 0.7    ; # want to estimate this
seq = rbinom(10,1, p)    ; # inputs
```

want to estimate p from the data

compare optimization of likelihood and log likelihood.

Notice: likelihood will tend to zero for large n!

Regression

Recall least squares fit:

$(x_1, y_1) \dots (x_n, y_n)$

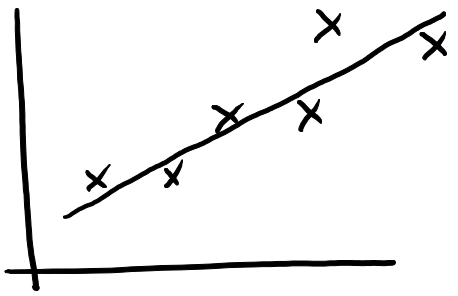
$y_{LS} = mx + b$ (linear model)

which minimizes:

$$R(m,b) = \sum_{i=1}^{n} (mx_i + b - y_i)$$

$$m = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$b = \frac{\sum x_i^2 \sum y_i - \sum x_i y_i \sum x_i}{n\sum x_i^2 - (\sum x_i)^2}$$



nonlinear least squares

$$f(x,a,b) = a\cos(bx) + b\sin(ax)$$

If the form of the non-linear model can be guessed, R provides a way to obtain the best fit coefficients a and b by solving an NLS (non-linear least squares problem).

$$R(a,b,x) = \sum \underbrace{| y_i - f(x_i,a,b) |^2}_{\text{non-linear function}}$$

Tests for the median (non-parametric stats)

The sign test ($n \leq 10$) is a simple non-parametric procedure for testing the hypothesis about the central tendency of a non-normal probability distribution.

Ex) Suppose the following carbon dioxide measurements are reported in standardized units / square area.

.78, .51, 3.79, .23, .77, .98, .96, .89

$n = 8$

$H_0: M = 1.00$
$H_1: M < 1.00$

} objective to determine if in an indefinitely large number of measurements, median < 1.

Test statistic: S = number of measurements less than 1.00, the null hypothesized median.

Suppose we wish to use $\alpha = .05$ level of significance.

Then, p-value = P(observing statistic as extreme or more extreme than one observed)

Rejection region: p-value $\leq \alpha = .05$

Here, $S = 7$ (7 of 8 measurements less than 1).

Let $X \sim$ Binomial(n,p) counting # of values below 1.

p-value $= P(X \geq 7)$

Note that if Ho is true, the binomial probability that a measurement lies below (or above) the median 1.00 is 0.5.

$$P(X \geq 7) = 1 - P(X \leq 6) = 1 - \text{dbinom}(6, 8, .5)$$

$$\approx 1 - .965 = .035 = \text{p-value}$$

Thus, the probability that at least 7 of 8 measurements would be less than 1.00 if the true median were 1.00 is only 0.035. Since p-value $< \alpha$ we reject the null hypothesis.

The true median is less than 1.00 at $\alpha = .05$ confidence level.

Note: for $H_0: M = M_0$
$H_1: M \neq M_0$

$$p\text{-value} = 2P(X \geq s)$$

where s = # of sample measurements $= \max(s_1, s_2)$

s_1 = # less than M_0

s_2 = # greater than M_0

Large-sample sign test for population median

$H_0: M = M_0$
$H_1: M \neq M_0$

$$z = \frac{(s - .5) - .5n}{.5\sqrt{n}}$$

Reject if $|z| > z_{\alpha/2}$

what is used is the normal approx to the binomial

simply normal approx for $X \sim B(n, p)$ with $p = 0.5$ and continuity correction

Ex) median time to failure for CD player $\alpha = .1$
claimed at 5,250 hrs. A sample of 20 CDs is used to test claim.

$H_0: M = 5250$
$H_1: M \neq 5250$

Suppose 14 of 20 samples exceeded 5,250 hrs.
6 were less than 5,250.

Test statistic:

$$z = \frac{(S-.5)-.5n}{.5\sqrt{n}} = \frac{13.5-10}{.5\sqrt{20}} = \frac{2.5}{2.236}$$

$$\approx 1.565$$

$z_{\alpha/2} = z_{.05} = 1.645$ reject H_0 if $|z_s| > z_{\alpha/2}$.

Thus, we cannot reject H_0.

$S = \max\{$measurements < 5250, measurements $> 5250\}$.