Maximum Likelihood Estimation

Many of the statistical tests for categorical data will use maximum likelihood estimation. Maximum likelihood estimation (MLE) is a method developed by R.A.Fisher (1950) for finding the best estimate of a population parameter from sample data. In statistical terms, the method maximizes the joint probability density function (pdf) with respect to some distribution. With independent observations, the joint probability of the distribution is a product function of the individual probabilities of events, so MLE finds the likelihood of the collection of observations from the sample. In other words, it computes the estimate of the population parameter value that is the optimal fit to the observed data.

MLE has a number of preferred statistical properties, including asymptotic consistency (approaches the parameter value with increasing sample size), efficiency (lower variance than other estimators), and parameterization invariance (estimates do not change when measurements or parameters are transformed in allowable ways). MLE can be seen as a more general method that encompasses ordinary least squares (OLS), where sample estimates of the population mean and regression parameters are equivalent for the two methods under regular conditions. MLE is applied more broadly across statistical applications, including binomial tests, contingency table analysis, logistic regression, and structural equation modeling.

Binomial Estimation

The goal of MLE in this context is to find the estimate of the population proportion, π, given the sample data (the set of responses on the binary variable with the sample size—the k success for n trials). The MLE of a proportion uses the binomial distribution function which we previously discussed (see the Probability, Proportions, and the Binomial Distribution handout).

$$P(Y = k; n, \pi) = \binom{n}{k} \pi^k (1-\pi)^{n-k}$$

This the statement of the binomial probability density function, and recall that the left-hand side $P(Y = k;n,\pi)$ refers to the probability of success for parameters n and π. The first term on the right-hand side can be ignored in the MLE process, because it is not needed to find the estimate of π. The remainder is termed the *kernel* of the likelihood function, and we find the likelihood $L(\pi)$ of this portion by using natural logs[1]

$$L(\pi) = \ln\left[\pi^k (1-\pi)^{n-k}\right] = k \ln(\pi) + (n-k)\ln(1-\pi)$$

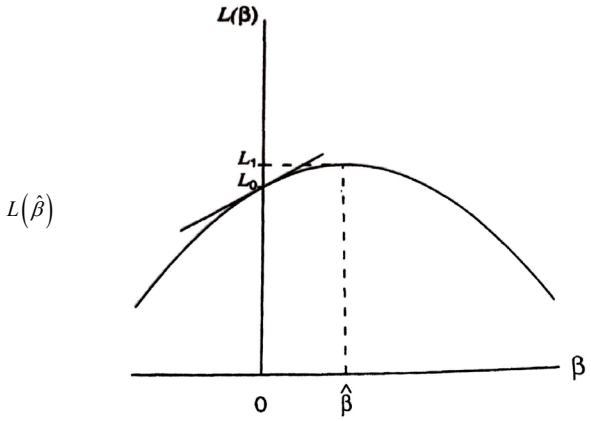
The second derivative is then used to solve for the parameter estimate after setting the first derivative of the above equation equal to zero and taking expectations. The result is the sample estimate p of parameter π given the data. By extension, we will also know the variance of the sample $p(1\text{-}p)$.

Iterative Process

For more complex problems, MLE is an iterative process in which initial (or "starting") values are used first. The computer then computes the likelihood function, which represents a lack of fit, for that set of parameter "guesses." On the next step, another set of parameter estimates are used and so on until there is a "response surface" that represents the likelihood values for all of the guesses. Each step is called an *iteration*. The idea is similar to the idea of ordinary least squares (OLS) in regression in which

[1] *exp*, the exponential function, and *ln*, the natural logarithm are opposites. The exponential function involves the constant with the value of 2.71828182845904 (roughly 2.72). When we take the exponential function of a number, we take 2.72 raised to the power of the number. So, exp(3) equals 2.72 cubed or $(2.72)^3 = 20.09$. The natural logarithm is the opposite of the *exp* function. If we take ln(20.09), we get the number 3. These are common mathematical functions on many calculators.

the squared errors or residuals are minimized to obtain the best fit of the regression line to the data and the regression coefficients.



p.12. Agresti, A.(2013). Categorical data analysis, third edition. New York: Wiley.

Tangent lines can be drawn (first derivative) for any particular point on the curve (as in the point L_0 in the figure above), and when the slope of the tangent line equals 0 (second derivative), the maximum of the curve of possible estimates is found, and this point corresponds to the optimal sample estimate of the parameter. The computer stops and generates values for the fit of the overall model and the parameter values.

Likelihood Ratio Test

The likelihood ratio test is a comparison of the maximum likelihood of the values under the null hypothesis and the maximum likelihood when the alternative hypothesis is allowed to be true (i.e., full null and alternative parameter space). In practice, we compare the fit of the expected values under the null hypothesis (L_0 = likelihood of E_i) to the fit of the observed data (L_1 = likelihood or O_i). Sometimes termed G^2, the likelihood ratio is -2 times the natural log of the ratio of observed to the expected likelihoods, or, equivalently, -2 times the difference of the natural log of the likelihoods.[2]

$$G^2 = -2\ln\left(\frac{L_0}{L_1}\right) = -2\left(\ln L_0 - \ln L_1\right)$$

Illustration with R

Instead of using all possible outcomes, let's make it a bit more concrete and say hypothetically that we have collected 25 polls asking respondents to choose Joseph Biden or Donald Trump. We can use the binomial distribution function to find the likelihood of the proportion of choosing one (we will use Biden here) out of the number of cases in each poll. The R code for setting the observed frequency of successes, k, the number of cases in each, n, and the observed probability, can be found below.

```
> #25 samples with observed frequencies of those favoring Biden, k = number success, n = total
> Biden <- c(22,40,11,33,27,30,25,25,20,19,44,27,28,30,34,24,28,29,31,19,24,29,33,32,25)
> k <- Biden
> n <- 50
> p = k/n
```

We can then use the `dbinom` function to obtain likelihood values, plot, and then optimize (find the maximum likelihood) with the `optimize` function can be

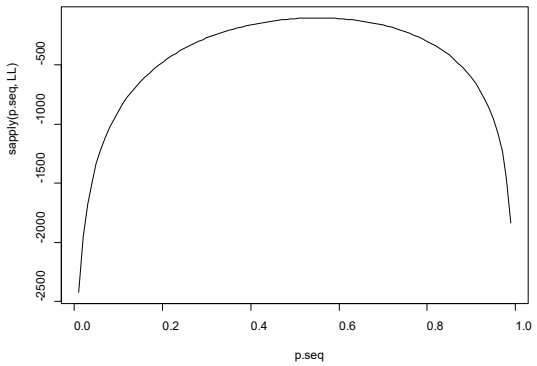
```
> # H1 loglikelihood
> LL<-function(p) sum(dbinom(k,n,p,log=TRUE))
```

[2] This is true in accordance with the quotient rule for logarithms.

```
> #plot LL:
> p.seq <- seq(0.01, 0.99, 0.01)
> plot(p.seq, sapply(p.seq, LL), type="l")
> #optimumimization of the observed data
> LL1 <- optimize(LL, lower=0, upper=1, maximum=TRUE)
> LL1
$maximum
[1] 0.5512004

$objective
[1] -101.7834
```

`$maximum` is the maximum likelihood estimate (expected value) of the proportion from the samples and `$objective` is the likelihood value. Notice the maximum of the curve is about .55.



For an individual sample, we would have a maximum likelihood value (e.g., $\ln L_0$ = -101.7834) and a maximized proportion estimate (e.g., p = .55) for the observed data that would be compared to the maximum likelihood for the null value with the proportion estimate of .5.

References and Further Reading

Agresti, A. (2013). Categorical data analysis, third edition. New York: Wiley. (pp. 9-13).
Eliason, S. R. (1993). *Maximum likelihood estimation: Logic and practice* (Vol. 96). Newbury Park, NJ: Sage Publications.
Fisher, R. A. (1950). Contributions to mathematical statistics. New York: Wiley.
Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, *47*(1), 90-100.