Available online at www.sciencedirect.com
SCIENCE DIRECT®
Statistical Methodology 2 (2005) 307–315
Statistical Methodology
Official Journal of the International Indian Statistical Association
www.elsevier.com/locate/stamet

Teachers' corner

An extended Minimum Description Length (MDL) penalty function for binomial proportions

Peter G. Bryant*

The Business School, University of Colorado at Denver and Health Sciences Center, P.O. Box 173364, Campus Box 165, Denver, CO, United States

Received 12 May 2005; received in revised form 26 August 2005; accepted 31 August 2005

Abstract

This article extends the Minimum Description Length (MDL) penalty function $B_2(n)$ suggested by P.G. Bryant and O.I. Cordero-Braña [Model selection using the minimum description length principle, The American Statistician 54 (2000) 257–268] to cover an arbitrary range of possible parameter values, and provides approximations for large sample sizes. The results are illustrated using data from a recent clinical trial (The Women's Health Initiative).

MSC: primary 62F99; secondary 62A99

Keywords: Minimum Description Length (MDL) Principle; Testing equality of two proportions

1. An introductory example

1.1. Background

As reported in [4], one arm of the Women's Health Initiative, the clinical trial of estrogen plus progestin, was terminated 3 years early in large measure because of an elevated rate of stroke among those women who took the estrogen plus progestin treatment compared to those who took a placebo. For a detailed report, including specific definitions and criteria, consult the article cited and the references therein. The following numbers are extracted from Table 2, page 2677 of that article. They are used here for illustrative purposes, and not to criticize the study

* Tel.: +1 303 556 5833; fax: +1 303 556 5899.
E-mail address: Peter.Bryant@cudenver.edu.

1572-3127/$ - see front matter
doi:10.1016/j.stamet.2005.08.002

or the decision to terminate it. The placebo group had $n_1 = 8102$ women enrolled, and of these $x_1 = 107$, or $p_1 = 1.3207\%$ had a stroke. The estrogen-plus-progestin group had $n_2 = 8506$ members, of whom 151, or $p_2 = 1.7752\%$ had a stroke.

1.2. Testing the difference between p_1 and p_2

To assess the statistical significance of the observed difference $p_1 - p_2 = .013207 - .017752 = -0.004545$ we may use an approximate (two-sided) z-test for the difference between the two population proportions. The standardized test statistic is

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)(1/n_1 + 1/n_2)}},$$

where $p = (x_1 + x_2)/(n_1 + n_2) = (107 + 151)/(8102 + 8506) = 1.5535\%$ is the pooled proportion of cases of stroke. This yields

$$\begin{aligned} z &= \frac{.013207 - .017752}{\sqrt{.015535(1 - .015535)(1/8102 + 1/8506)}} \\ &= -2.367, \end{aligned}$$

which would be statistically significant at, say, level $\alpha = .05$, though not at level $\alpha = .01$.

The result was widely reported in the press, and may have influenced women and their physicians to consider stopping the use of the estrogen-plus-progestin treatment, though the risk of stroke while taking the treatment is still relatively small and the treatment does have benefits.

1.3. An approach using the Minimum Description Length (MDL) Principle

We may also assess the observed difference using the Minimum Description Length (MDL) Principle. According to (one version of) MDL, the two proportions p_1 and p_2 would be considered convincingly different if

$$nH(p) + \ln B_2(n) > n_1 H(p_1) + \ln B_2(n_1) + n_2 H(p_2) + \ln B_2(n_2), \tag{1}$$

where $H(t) = -t \ln t - (1-t)\ln(1-t)$ and

$$B_2(n) = \sum_{x=0}^{n} A_{x,n} \tag{2}$$

with

$$A_{x,n} = \binom{n}{x}\left(\frac{x}{n}\right)^x \left(\frac{n-x}{n}\right)^{n-x} \tag{3}$$

$$= \left(\frac{n!}{n^n}\right)\left(\frac{x^x}{x!}\right)\left(\frac{(n-x)^{n-x}}{(n-x)!}\right). \tag{4}$$

$B_2(n)$ is the MDL penalty function described in [3].

For the Women's Health Initiative example above, these calculations yield

$$\begin{aligned} nH(p) + \ln B_2(n) &= 16608 \times H\left(\frac{258}{16608}\right) + \ln B_2(16608) \\ &= 16608 \times .08011037 + 5.088735 \\ &= 1335.562 \end{aligned}$$