

CSE 291: Statistical Learning

Penultimate Assignment

This assignment is due in class on Tuesday March 1, 2005. All the guidelines from previous assignments still apply. See also the January 25 feedback on Assignment 1.

Please use <http://www.quicktopic.com/29/H/t3sgTnDZkMUqp> to ask questions about the problems below.

(1) [Silvey, Problem 7.3.] An experiment has $N + 1$ possible outcomes z_0 through z_N . The null hypothesis H_0 assigns probabilities to these as follows: $P(z_0) = 1/2$ and $P(z_n) = 1/2N$ for $n \geq 1$. The alternative H_1 assigns even higher probability to z_0 , namely $P(z_0) = 1 - 1/N$, and does not specify the individual probabilities of z_1 to z_n .

Consider the likelihood ratio test of H_0 versus H_1 with size $1/2$, based on a single observation. Show that this test accepts H_0 if and only if the observation is z_0 .

What is the power function of this test? Intuitively, is it a sensible test?

Reminder: The power function $\alpha(\theta)$ of a hypothesis test is the probability of rejecting H_0 when θ is true. We want $\alpha(\theta)$ to be low for all $\theta \in H_0$, and high for all $\theta \notin H_0$.

(2) Like most textbooks on statistics, the NIST Engineering Statistics Handbook says “For the χ -square approximation to be valid, the expected frequency should be at least 5” [in each cell]. Similarly, Prof. Carl Schwarz of Simon Fraser University wrote the following in lecture notes:

“The test statistic can be computed in several ways:

- Pearson χ -square which is formed by comparing the observed and expected counts when the null hypothesis is true.
- Likelihood ratio χ -square which is formed by a weighted average of the ratio of the observed and expected counts.
- [...]

The first two test statistics often are very similar and there is no objective way to choose between them. It can be shown that in large samples the two are equivalent. However, both test statistics assume that the counts in the contingency table are reasonably large. An often quoted rule of thumb is that the expected count in each cell of the contingency table should be at least 5.”

(a) Investigate experimentally how closely the LRT statistic and the χ -squared statistic follow χ -squared distributions, when the total sample size is small. Generate figures that illustrate your conclusions persuasively.

You may use Section 4.3 and Chapter 5 of *Finding Structure in Text, Genome, and Other Symbolic Sequences* by Ted E. Dunning (Ph.D. thesis, University of Sheffield, 1998) for inspiration, especially Figure 5.6. This text is available at <http://www-cse.ucsd.edu/users/elkan/291/dunningLRT.pdf>. However, you do not need to work through all Dunning's equations.

(b) Investigate experimentally the claim that the LRT statistic “assume[s] that the counts in the contingency table are reasonably large.” That is, look into how closely the LRT statistic and the χ -squared statistic follow χ -squared distributions when the total sample size is large, but the number of observations in one or more cells is small.

(c) Based on your results, explain whether Professor Schwarz is right that “there is no objective way to choose between them.”

Can you suggest a guideline that is better than the “less than 5” guideline for when neither test should be used, or one can be used but the other should not be, or both can be used?

(3) (a) Derive a likelihood ratio test for including or excluding sets of predictors in a linear regression model.

(b) Run simulations to investigate the power function of your test, using synthetic datasets where you know the true answer.

(c) Run similar simulations to investigate the power function of the usual F -test. Draw conclusions.

(4) (a) For linear regression problems with just one predictor, derive simple formulas for the slope and intercept that do not involve matrix computations.

(b) Each year, a research team at Colorado State University predicts how many hurricanes (of three different intensity levels) will strike the eastern U.S. the next year. The predicted and actual numbers are attached. Explain how to use your answers to Problem 3 and part (a) to decide whether or not the forecasts have any value.

(c) Apply your method from part (b) to the given data, once for each of the three storm types. Interpret your results.