

Re: Chi-square for binomial samples

Source: <http://sci.tech-archive.net/Archive/sci.stat.math/2006-06/msg00002.html>

- *From:* David Winsemius <doe_snot@xxxxxxxxxxxxx>
 - *Date:* Wed, 31 May 2006 17:50:19 -0500
-

Michael McLaughlin <mmclaughlin1@xxxxxxx> wrote in
[Xuofg.24929\\$ZW3.3160@dukeread04](mailto:Xuofg.24929$ZW3.3160@dukeread04):">news:[Xuofg.24929\\$ZW3.3160@dukeread04](mailto:Xuofg.24929$ZW3.3160@dukeread04):

Scenario:

N experimenters take samples from a common population, effectively infinite, BUT these samples are of various sizes. They each look for a given (fixed) attribute of interest and record their frequency of success.

Were their samples all of the same size, then every statistics book ever written would discuss how to compute the associated chi-square statistic.

Most chi-square statistics that I have seen (and there are many) assume variable numbers in each stratum. Can you give us some examples?

Moreover, computing a maximum-likelihood value for p is still straightforward since every term in the log-likelihood is well-defined.

Question:

Is there an accepted method to compute chi-square under the conditions stated — where p is assumed known but sample size is variable? Is this even a sensible question?

The chi-square is just the sum of squared deviations from the expected. Usual basis for expected is row-sum \times column-sum/total. Are you proposing to substitute a known p (times the row total presumably) for that number? If so, you may want to investigate GLMs with offsets.

The analogous question could be asked wrt the beta-binomial distribution as well. There, the scenario described is, in fact, the

Re: Chi-square for binomial samples

norm. The question could be asked again, a fortiori, wrt the hypergeometric distribution (two parameters nominally fixed).

It sounds as though you are asking for a test of equal proportions among k multiple samples. Results (successes, failures and associated marginals) would conventionally be displayed in a 2 column x k row table with row and column sums. The X -squared statistic would be compared to chi-squared distribution with $(k-1)$ degrees of freedom. The row totals do not need to be equal. The usual condition for validity is that the number of cells with counts below 5 is fewer than 20%.

Your added information about "p being known" may be a ringer. What sort of global hypothesis are you thinking of testing when the proportion of successes is already given?

--

David Winsemius

.