Re: significance in contingency table

Source: http://sci.tech−archive.net/Archive/sci.stat.math/2004−10/0238.html

---

From: François Charton (*fcharton_at_carthage.fr*)
Date: 10/14/04

Date: Thu, 14 Oct 2004 22:14:07 +0200

"Aleks Jakulin" wrote:
>
> *Actually, both your approaches are sensible. The first approach is*
> *trickier to implement, but the idea is that you compute the test*
> *statistic just in the cell of interest, not for the whole table. But*
> *you have to derive the corresponding distribution of the statistic*
> *under the null, and the result will depend on the number of columns*
> *and rows, which is not necessarily intuitive. We are assuming that*
> *we're trying to examine the behavior of predicting I+J in a larger*
> *model.*
>

Actually, the objective is to be able to help detect the main "features" of a table. My data comes from large sample surveys, and I want to be able to help non specialists to read such tables. For instance, if I had a table splitting a population by age groups and regions, I would like to be able to detect that some class is overrepresented in some region, or that some region has this or that specificities in term of the age of its residents...

I understand the need of finding the law of the statistics under the null hypothesis (independence).

The first approach seems the most natural, for the law of the statistics, the formula hints at a Chi−square with one degree of freedom: the underlying data can be supposed to be binomial, for large samples, it converges to a normal, and the statistics seems to be homogeneous with its square (once you center it and normalise it).

Now, and this is the heart of the problem, if I try the second approach, and apply a chi−square test, I get a larger statistics (summing the previous value, plus the contributions of the three other cases), but still compare it to a Chi−square law with 1 degree of freedom...

So, the second test would fail in many cases where the first succeeds, which seems strange. What am I missing?

Francois