

Re: Cluster analysis for beginners

Source: <http://sci.tech-archive.net/Archive/sci.stat.math/2007-03/msg00736.html>

- *From:* Jerry Dallal <gdallal@xxxxxxxxxxxxxxxxxxxxxxxx>
 - *Date:* Thu, 29 Mar 2007 19:36:59 -0400
-

illywhacker wrote:

On Mar 29, 4:38 pm, David Winsemius <doe_s...@xxxxxxxxxxxx> wrote:

Sidney <milan_y...@xxxxxxx> wrote
innews:24466740.1175159875339.JavaMail.jakarta@xxxxxxxxxxxxxxxxxxxxxxxx:

Assume you have 5000 proteins that are ordered by their molecular weight from 1000 Daltons to 100000 Daltons (the numbers don't matter). If you now find that a certain motif (e.g. a specific phosphorylation motif) which is only found within a certain molecular weight range, e.g. only between 77000-81000 Daltons, how do you determine if this 'clustering' is significant? At this point I have no idea what to do and where to start at. Your input is very much appreciated. Thanks a lot in advance. Sidney

I did see illywacker's reply, but I disagree. I thought your scientific question was reasonably clear for one thing. If your null hypothesis is that there is no association between MW and presence of the motif, you could start by arranging the proteins in deciles of weight and testing for uniformity of motif-ication in a multinomial model with 9 degrees of freedom. Unless most of your proteins are in that specified range above, with 5000 data points it seems reasonably clear that you will get a 'significant' result using that approach. I would argue that such a test does not represent one with "strong assumptions".

1) Classical hypothesis testing is fatally flawed. No well-defined alternative is specified, and the probability of the data is not calculated. Rather the probability of a set of unobserved data points is

Re: Cluster analysis for beginners

calculated. As Jeffreys famously put it: "A hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred". There is a mass of literature on this.

This is a "joke", of course, that results from thinking of P values as posterior probabilities. If P values are thought of in terms of fixed level tests, Jeffreys' comment makes no sense.

Asking whether the clusters are "significant" is too vague to answer. I suspect what the OP meant was whether the clusters are "remarkable". In order to answer that, one must first address the question, "remarkable compared to what?" Possible answers might be, "Compared to a uniform distribution of proteins" or "Compared to a unimodal distribution of proteins with a peak of XXX Daltons..." The distribution theory for clusters is remarkably difficult in general. However, this is one place where good results could be obtained through simulation. That is, construct a measure of "remarkableness", generate a bunch of samples that follow the "compared to" distribution, assess the distribution of "remarkableness".

2) However, classical hypothesis testing does in fact, despite appearances, have an underlying set of alternatives in any given (well-defined) case. These usually correspond to very simple models that have nothing to do with the situation in hand.

3) In this case, you use the word 'association', which is not defined. How can you test for the absence of something when you do not know what it is? What does the test actually tell you?

4) You assume that nothing is known about the possible 'associations' a priori, which may or may not be true.

5) Why deciles? How does the result change if you use n-iles? This relates to prior knowledge of the 'smoothness' of the 'association', but you do not discuss it.

Off the top of my head, and without analysing this particular problem, that will do for starters.

illywhacker;

Re: Cluster analysis for beginners