Re: Shotgun statistics

Source: http://sci.tech−archive.net/Archive/sci.stat.math/2005−01/0583.html

---

RossClement_at_gmail.com
Date: 01/26/05

```
Date: 26 Jan 2005 04:03:42 −0800
```

Richard Ulrich wrote:
> On 24 Jan 2005 04:33:17 −0800, "RossClement@gmail.com"
> <RossClement@gmail.com> wrote:
>
> > Hi. I've been thinking about the problem of using a confidence
level of
> > 95% for deciding whether to reject a null hypothesis. I've always
> > assumed that this would mean that 5% of experiments that should
show no
> > effect, would show an effect by chance. However, if I try to
calculate
> > the number of hypotheses that need to be considered before we have
a
> > >50% of chance of finding a spurious effect, I get the following.
> >
> > (i) Given a 95% confidence level for rejecting the null hypothesis,
and
>
> I try not to talk about these, because it gets confusing so easily.
> Shouldn't that be, "at least 95% confidence of accepting the null"?

Hmmmm..... I must admit that I'm thinking in a "Monte−Carlo style"
here. I think that it's best that I ask the question again, using a
simple invented example.

Let's assume that we build a model of the null hypothesis. Just for the
sake of argument, assume that the null hypothesis is that we have two
variables A and B, both of which are normally distributed and
independent. We have a random sample of 50 (a,b) pairs. We want to know
if the correlation we measure between A and B in this sample is
sufficiently high (or low) to reject the hypothesis of independence
between A and B.

If we were to evaluate this by Monte−Carlo simulation, and a 95%
confidence interval, then we could implement the null hypothesis as a
computer program, and generate millions of random 50 (a,b) sets. We
could then calculate the correlation between A and B in each of these

sets, which would give us the expected distribution of correlations if the null hypothesis was true (call this d). If we then construct a symmetric 95% confidence interval around the mean of d, then we can see if the correlation from our original set of data is extreme enough so that it is outside that confidence interval. If within the confidence interval, we don't reject the null hypothesis. If outside the confidence interval, we reject the null hypothesis and accept the alternate hypothesis that A and B are correlated.

Note: I do realise that there are better significance tests for correlation.

Now, given the method we have above, even if the original sample was generated by the null hypothesis, there is a 5% chance that it would be rejected by the above significance test, as the confidence interval is set so that 5% of the samples generated will lead to us rejecting the null hypothesis with 95% confidence, no?

Given that situation to work on, I'll try to phrase my question better using A and B as an example.

Let's assume that the null hypothesis is true, A and B are both normally distributed and independent. Then, how many randomly selected samples of size 50 do we need to select before the probability of at least one of them showing a "statistically significant" correlation between A and B. The chance of one such sample being independent is 5% or 0.05. Hence the chance of us correctly failing to reject the null hypothesis given a sample is 0.95. If we're selecting N such samples (each sample being 50 pairs), what is the smallest value of N such that the probability that at least one pair has a "statistically significant" correlation? The probability that no pairs have a statistically significant correlation is 0.95^N, and the smallest N for which 0.95^N < 0.5 is N=14.

In the case above, the chance of falsely rejecting the null hypothesis is 5% or 0.05 as we assume that the null hypothesis is true, and the Monte−Carlo simulation is accurate. As Rich points out, in real life we're rejecting the null hypothesis with *at least* 95% confidence, and hence the probability of falsely accepting the alternate hypothesis is < 5%, not = 5%.

> > *a proper alternate hypotheis which is the true negation of the null*
> > *hypothesis, then I assume that the chance of incorrectly rejecting*
the
> > *null hypothesis is 0.5,*
>
> *Suddenly, you "assume" the alpha is 0.5? or was that a typo?*

That is a typo. It should be 0.05.

> > and the probability of correctly failing to
> > reject the null hypothesis is 0.95.
>
> Was (i)" supposed to be an easy re−statement of definitions
> which got messed up, or were trying to assert things here?

I thought it was best to start the question again rather than try and
fix the previous version. I hope my new question doesn't have the same
problems.

> > (ii) Lets assume that we have a single set of data, and a large
number
> > of null and alternate hypotheses that can be investigated (e.g. the
> > astrobank dataset used for investigating astrology). Lets also
assume
> > that all of the alternate hypotheses are spurious (i.e., in this
> > example, that "astrology is bunkum"). If we keep on choosing
hypotheses
> > and investiging their statistical significance, then the
probability
> > will get higher and higher that we will find some spurious
hypothesis
> > that "gets lucky" and comes out significant.
>
> This is why we speak of experiment−wise error. And corrections.

I've had an email response (thanks!) which recommended that I read up
on Bonferroni corrections and Tukey's HSD. I presume that your sentence
here is talking of similar things.

> > (iii) If we view the probability that we have at least one such
> > hypothesis coming up out of N, then this is one minus the
probability
> > that no such hypotheses are found. Assuming that testing hypotheses
are
> > independent random events, then the probability of this is:
> >
> > (0.95)^N
> >
> > (iv) A quick calculator check shows that this is < 0.5 for N>=14.
> > Hence, if I use a 95% confidence level, and choose 14 or more such
> > alt/null hypothesis pairs, the probability that I get at least one
> > improper reject of the null hypothesis is better than even.
> >
> > Is my reasoning and calculation correct?
> > Note: this is not a homework problem.
>
> There's an assumption in here, that the tests
> are full power (5%). Consider testing for "fair coins"
> with sets of 4 flips −− No rejections.

Hmmm.... But for normally distributed variables such as A, B, with a sufficient sample size, such as 50, or 200 values per sample, as described above, the tests would be full power, would they not?

This question isn't too important to me. I have been inspired by a paper that was mentioned in a book I read where the likelihood of invalid results in medical studies was evaluated using simulation. I'm looking into performing a similar simulation to look at problems that could occur when methods for authorship attribution are evaluated on small, and single, data sets. In that case, I won't be assuming that there is a fixed, or even known, probability of an invalid conclusion, but will evaluate the method variations on large numbers of small sized data sets, so that I can empirically estimate such things as the probability that a "statistically significant" improvement found on a small data set represents a true improvement given more robust testing, etc. I'm talking to a colleague in our Maths&Stats department to see if he's interested in collaborating on this.

And, to be honest, if I can gain more skills in this area, I'd love to have a crack that that "astrobank" astrology research databank :−)
Cheers,

Ross−c