

Re: Pattern matching with statistics

Source: <http://sci.tech-archive.net/Archive/sci.stat.math/2006-02/msg00290.html>

- *From:* junk5@xxxxxxxxxxxxxxxxxxxx
 - *Date:* 24 Feb 2006 02:39:35 -0800
-

If I have a set of random data, which could be the result of rolling a 6 sided die 1,000 times, and the die is favored to rolls 2 numbers more often than the others, how do I analyze the data to determine which numbers it favors without knowing in advance that it favors any of the numbers?

A simple approach to this involves performing a χ^2 hypothesis test. Your null hypothesis would be that the distribution underlying your sample is the same as the 'fair' distribution. So, in the case where you have samples drawn by throwing a single 6-sided dice, your null hypothesis would be that there is no difference between the distribution underlying the samples and the discrete uniform distribution for 6 values (i.e. the numbers 1-6 each have probability $1/6$). By choosing an appropriate significance level, you can then work out if it is likely or unlikely that the two distributions are the same.

So this will tell you if it's likely that some numbers are favoured over others, but it won't tell you which ones. You cannot just look for the most common number(s) in the sample, as you need to know which ones are statistically significantly more common. You might want to think carefully about whether this is what you want or if the answer obtained from χ^2 will solve your problem.

Normally the gamblers fallacy ... [snip]

I think you're misusing this term here. The Gamblers Fallacy is the phenomenon that you see in gamblers where they think that an event that in reality has a fixed probability is more or less likely if that event has recently occurred. In you example, you seem to be talking about "sampling without replacement". In this kind of sampling, it is events that have not happened yet that become more probable (and eventually inevitable when there is only one item left to be sampled).

Lets say for this example that the machine is programmed to never

Re: Pattern matching with statistics

produce the same number twice, until it has randomly generated every other possible number. Is there a way to predict this is happening by looking at the data?

I'm not sure exactly what you mean here. Here's some possibilities:

i) You have a sample of data, you know the order in which each datum was generated, you know the data should have been generated by sampling without replacement and you want to know if the data was indeed generated in that way.

ii) You have a sample of data, you *don't* know the order in which each datum was generated, you know the data should have been generated by sampling without replacement and you want to know if the data was indeed generated in that way.

In case i) it should be very simple to rule out sampling without replacement if you know the range of values that can be sampled and have a large sample. If you see the same number twice before all possible numbers have been sampled, then sampling without replacement cannot be happening. You will need to think carefully about the case where you only have a small sample though.

In case ii) you may also be able to rule out sampling without replacement by considering the probability of a given value occurring in the size of sample that you have.

.