Try, try, and try again

SIR — Bill Amos warns that the accepted standard approach to statistics can easily lead to erroneous results (*Nature* **379**, 484: 1996). This is because, over the whole world, many scientists may be studying the same phenomenon or closely related phenomena, and several different scientists may perform the same experiment or a similar experiment unknown to each other. Even if the experiment should show no correlation at all, as more scientists try, the chances of one of them producing a 'statistically significant' (and erroneous) result quickly rise.

Although Amos states the case well, he cites numbers of investigators that might seem large enough to make such happenings unlikely or even unrealistic. Actually, the way science operates plays so directly into this statistical trap that large numbers of unwitting participants are not required. In science, failures are very rarely mentioned. Thus, if an experiment is a logical next step, the idea will naturally occur to scientists in its field. One after another will pursue it until one of them succeeds and is published. If after publication another scientist tries the same experiment and fails to replicate it, that typically will go unreported because such a report is very difficult to get past the referees and might tarnish the author's reputation in the process. Usually it is only when a result proves a major stumbling-block to continuing research that any real effort is made to unseat it. Much more often, if a scientist finds that a result cannot be replicated, then a way is found to sidestep it.

The most important effect of most studies (erroneous or not) is that they influence future work, directly or indirectly. As scientists ponder what to do next, the erroneous study will influence some scientists as they design experiments in parallel courses and other scientists as they design experiments in other areas that seem logically related. If more than thirteen experimental trials which should show no correlation are influenced by the erroneous study, the chances are better than even that at least one will show $P<0.05$. The cycle continues. And, those experiments that do not show significant results will go unreported and so will still seem a logical course to other scientists, waiting to be tried again. Some of these re-tried experiments show statistical significance....

This becomes much more alarming when it is realized that this description of scientific practice is modelled very closely by a now classic combinatorial probability game in which a black ball and a white ball are placed in an urn (R. M. May, *Nature* **262**, 646; 1976) If one draws a white ball, then that ball is replaced and another white ball is added to the urn for future draws. If one

draws a black ball, it is replaced and a black ball added to the urn. The game continues with successive draws and is scored by taking the ratio of white balls to black balls in the urn on successive runs. Thus the result of each draw or 'experiment' influences the result of future experiments.

If one tries this game, one finds that the ratio quickly stabilizes to a steady value of 0.674... and is quite deterministic in settling to that result. One might think this result strange, but it is easily explained when one understands that the surface texture of the two colours of ball are slightly different, probably influencing the draw. With this simple game, however, one has the distinct advantage of being able to start again from the beginning, with one black and one white ball. The ratio will again quickly stabilize in a quite deterministic manner. It will, however, stabilize to a completely different value, say 0.228... The surface texture explanation is meaningless. Each time the game is played, it appears as deterministic as any chain of scientific studies, but each time the final deterministic-appearing result is different.

This same mechanism, caused by the inevitable interaction of statistics and standard scientific practice, will also lead to deterministic-appearing chains of scientific development which are quite false. There do not have to be large numbers of scientists doing the same experiment for this to happen.

David Dunthorn
C F Systems,
908 West Outer Drive,
Oak Ridge, Tennessee 37830, USA

SIR — Amos illustrates in a fictitious scenario how chance might lead to reporting of a non-existent 'finding' (type 1 error). He appears to have used one-tailed tests throughout his illustration and to have overlooked the probability of a similar distribution of his 50 'significant' (at $P <0.05$) findings among 1,000 trials in both tails of the distribution. If the spurious 'findings' were in truth nonexistent, the 1,000 trials would yield only some 25 'significant' positive reports balanced by some 25 'significant' negative reports (and 950 nonsignificant outcomes supporting the 'null hypothesis'). Publication of an unbiased mix of 'significant negatives' and 'significant positives' would end the illustration there.

Although truly 'negative' findings (not 'null' findings) may be as important to science as 'positive' findings, scientists often seek trends in one direction, which one depending on the field of enquiry and the current consensus paradigm. Publication bias may distort the fair finding of 25 positive trials and 25 negative trials by selective-

ly reporting more of those that show an association in one direction (perhaps the 'positive'). It is not the play of chance that inhibits scientific advance so much as the play of prior prejudice, seeking significant findings in only one tail of the distribution.
Robert West
University of Wales
  College of Medicine,
Heath Park,Cardiff CF4 4XN, UK

Legal pill

SIR — In their Commentary article "Why Japan ought to legalize the pill" (*Nature* **379**, 579–580; 1996), Maruyama *et al.* say that "Japan is the only industrialized country, except for the Republic of Ireland, where steroid oral contraceptives are illegal". That is not correct. Steroid oral contraceptives have been available for prescription in Ireland for more than 15 years, and the range of such products available is comparable to that in any other European country.
John G. Kelly
Irish Medicines Board,
63-64 Adelaide Road,
Dublin 2, Ireland

In the ears of the beholder?

SIR — As Hunt and Balsan have pointed out[1] musicians have many beliefs about violin tone, but it was established long ago that hearing cannot distinguish between the bowed sounds of old and new violins[2].

High damping is desirable in spruce[3] and in strings[4]. Wolf notes are due to low damping. When a violin correctly set up at 60–70 per cent relative humidity is exposed to the dry air of a centrally heated atmosphere, it changes shape and, for example, the gap between strings and fingerboard may be reduced so as to make it unplayable; in comparison, any changes in the physical properties of the wood are trivial.

If playing or age changes wood or other vital components such as glue, a skilled player compensates for the properties of each individual instrument, and there is no way in which the effect of any property can be detected by a listener[5]. Most of the beliefs about violin tone appear to be subjective.
James Beament
Queens' College,
Cambridge CB3 9ET, UK

1. Hunt, D. G. & Balsan, E. *Nature* **379**, 681 (1996).
2. Saunders, F. A. *J. Frank. Inst.* **229**, 1–20 (1940).
3. Cremer, L. *The Physics of the Violin* (MIT Press, 1985).
4. Pickering, N. C. *The Bowed String* (Amerion, New York, 1989).
5. Beament, J. *The Violin Explained* (Oxford University Press, in the press).