



© 1997–2004, Millennium Mathematics Project, University of Cambridge.

Permission is granted to print and copy this page on paper for non-commercial use. For other uses, including electronic redistribution, please contact us.

September 2003

Regulars

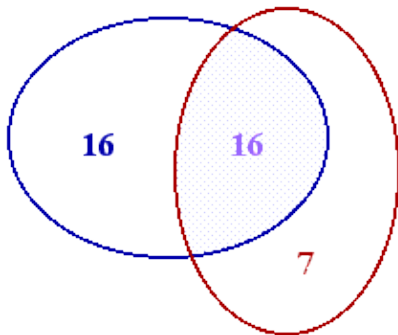


Outer space: Independence Day

by John D. Barrow



Independence Day, July 4th, 1977 is a day I remember well. Besides being one of the hottest days in England for many years, it was the day of my D.Phil. thesis examination in Oxford. Independence, albeit of a slightly different sort, turned out to be of some importance because the first question the examiners asked me wasn't about cosmology, the subject of the thesis, at all. It was about statistics. One of the examiners had found 32 typographical errors in the thesis (these were the days before word-processors and spell-checkers). The other had found 23. The question was: how many more might there be which neither of them had found?

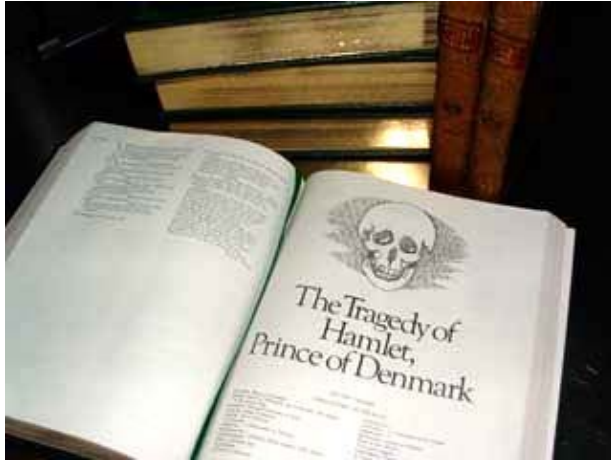


After a bit of checking it turned out that 16 of the mistakes had been found by both of the examiners. Surprisingly, knowing this information means you can give an answer so long as you assume that the two examiners work independently of each other, so that the chance of one finding a mistake is not affected by whether or not the other examiner finds a mistake. Using this assumption of statistical independence, can you show how I was able to say that we expected there to be only another 7 errors that both examiners missed?

Outer space: Independence Day

The general formula is a useful one to know. What is the number of unbound errors if one reader finds A , the other finds B , and they find C of those in common? If you like algebra you could work out the generalisation to 3, 4, or N proof-readers.

This type of argument has been used in many practical situations. Suppose different oil prospectors search independently and find different oil pockets: how many might lie unbound? Or suppose that ecologists want to know how many animal or bird species might be in a region of forest if several observers do a 24-hour census?



A similar type of problem arises in literary analysis. In 1976, two Stanford statisticians, Bradley Efron and Ronald Thisted (see *Biometrika* 63, 435–47 (1976)), used a similar approach to estimate the size of William Shakespeare's vocabulary by investigating the number of different words used in his works, taking into account multiple usages. Shakespeare wrote about 900,000 words in total. Of these, he uses 31,534 different words, of which 14,376 appear only once, 4,343 appear only twice and 2,292 appear only three times. They predict that Shakespeare knew at least 35,000 words that are not used in his works: he probably had a total vocabulary of about 66,500 words. Not bad!

Did you manage to answer the puzzle posed in [Outer space: A sense of balance](#)? If not, you can [find the answer here](#)!



Plus is part of the family of activities in the Millennium Mathematics Project, which also includes the [NRICH](#) and [MOTIVATE](#) sites.