

Even if the test had been based on 1000 randomly ~~chosen~~ chosen cards, so that 278 were required of each red suit, a pack with exactly ~~25~~ 25% of each ~~red suit~~ ~~red suit~~ provide an apparent

TP Even if the test had been based on 1000 randomly chosen cards, so that 278 were required of each red suit, a pack with exactly 25% of each, would ~~not~~ provide an apparent disproof with a frequency of <sup>only</sup> about 38 trials in a million.

Even if the test had been based on 1000 randomly chosen cards so that 278 were required of each red suit, a pack with ~~exactly~~ exactly 25% of ~~each~~ each would provide an apparent disproof with a frequency of only about 38 trials in a million

8 / single

Hearts could be excluded at a reasonable level of significance, if 34 Hearts appeared in 100 chosen. If the true proportion were 25%, the probability of observing 34 or more is found to be 2.759%. However, the composite hypothesis is disproved only if it is demonstrable that the proportion of Diamonds also is more than 25%, and this requires that the sample should contain at least 34 Diamonds. It is not difficult to anticipate that both these conditions together will be fulfilled very rarely, even in the case in which both Hearts and Diamonds contribute a full quarter to the material sampled. In fact, an apparent disproof of the composite hypothesis at the moderate level of significance chosen would in this case be obtained in less than 34 trials in a million.

It is, of course, no inconvenience that the frequency of rejecting the hypothesis in some cases when it is true should be low, but the calculation indicates also that even if both suits did really occupy somewhat more than 25% of the material sampled, it would not be easy to demonstrate this fact, even at a moderate level of significance, if neither of them were greatly in excess. Sufficiently large samples could indeed make such a demonstration probable; but the frequency of attaining a significant sample in the limiting case where both red suits have exactly 25%, is always less than the square of the fraction measuring the level of significance.

In scientific work it is necessary to be able to assess the strength of the evidence that a particular hypothesis, simple or composite, appears to be untenable. The example has been chosen to show that strong evidence may sometimes be hard to obtain. Warnings that the strength of the evidence