The consolation of the marathon experimenter

The physicist and the chemist may be able to perform experiments using homogenous material and even the biologist can use genetically pure mice but the clinical scientist, like the agronomist, has to accept variability. Just as plots of land differ in fertility, patients suffering from a given disease may vary considerably in symptoms and eventual course of the disease. The only guarantee that randomisation appears to offer when constructing treatment groups is that of an irrelevant long run: 'the consolation of the marathon experimenter'. On average, over all possible randomisations, the groups will be equal.

Is this a comfort? If we are 35 000 ft above mid-Atlantic, all four engines are on fire and the captain has had a heart-attack, can we console ourselves with the thought that on average air travel is very safe? No. Nor would R. A. Fisher expect us to. We would be faced with what he referred to as 'a recognisable subset', a rather unpleasant one in this case. Fisher never intended that randomisation should exclude your taking other precautions as regards design. As already explained, he suggested, in fact, that you should balance designs for what you believed to be relevant and then randomise subject to these restrictions. Fisher also pioneered the use of models that allowed you to adjust for any imbalance after the event.

Such models will be mentioned in subsequent chapters but a simple illustration of the way they work can be given by considering the case of sex. Suppose we have randomised but *not* balanced by sex and we notice that the split is 80 males and 20 females in the active treatment group and 70 males and 30 females in the placebo group. Since, as we have already noted, other things being equal, females tend to have a lower FEV_1 than males, if we do not take account of this, the results will be biased in favour of the active treatment. We can deal with this quite simply, however, by calculating the difference between the mean of the 80 males in the treatment group and the mean of the 70 males in the placebo group, and the corresponding difference for females. Each of these two differences is an estimate of the treatment effect for the corresponding sex. It is often plausible, given a suitable choice of measurement scale, to believe that this effect does not differ too much from one group to another. (This does not require that males have similar FEV_1 values to females but simply that the *difference* that a drug makes to this measure will be similar for males and females.) It thus becomes useful to form an average of these two treatment estimates. There is a further complication whose details we cannot discuss