A major object of the ‘Measurement in Medicine’ course given to first year pre-clinical medical students at The London Hospital Medical College is to enable them to understand the need for proper experimental design; they are also introduced to elementary methods of statistical analysis. The students receive a short series of lectures, and are divided alphabetically into eight groups for tutorials, two groups of 10 students being assigned to each of four tutors. The course itself is regarded as a research tool in teaching methods and has been the subject of previous reports 1–3.

The two groups assigned to me carried out during their tutorial sessions an experiment upon their abilities to distinguish the taste of different beers. The main question, as a result of in-class discussion, was whether individuals could discriminate between light and bitter ales solely by taste, and also name them correctly. The experimental method used was formulated during two tutorials before the project was actually carried out. The question whether one can ‘tell Stork margarine from butter’ is similar, and has been the subject of a previous (unpublished) experiment by another tutor.

Methods

Two groups of nine students (14 men and four women) were tested on separate occasions, one from each group being absent on the day of the experiment. So that taste should be the only means of discrimination available, environmental conditions that might have interfered were, as far as possible, standardized. All testing for each group was carried out during a single one hour session, the subjects were blindfolded to eliminate visual aid, the two types of beer compared (bottled light and keg bitter ale from the same brewer) were at the same temperature, and exactly one-eighth of a pint was given in identical plastic beakers on each occasion. Only one of two answers was permitted- ‘light’ or ‘bitter’. There was thus a one in two chance of being correct by guessing on each test, a one in four chance of being correct on two tests, and so forth. With five consecutive tastings the probability of guessing all correctly is small enough, one in 32, to reject at the conventional 5% level the null hypothesis that they are unable to discern light from bitter. Each student, therefore, tasted five beakers at intervals of six or seven minutes, and then handed in a completed response record. The tasting was done in a separate room from the remainder of the group with only the experimental assistant present, and no answers were given before the completion of all five tastings.

Results

The order in which the beers were given to each student and those which they correctly identified are shown in Table 1. All figures quoted from here on are derived from this table.

Only two of the 18 students (Nos 1 and 17) gave statistically significant evidence of rejection of the null hypothesis. One of these was a girl; the other said before the experiment took place that he thought he could discriminate the two tastes though perhaps not label them correctly, but in fact he did both. One other student had four correct responses out of five. Ten made more incorrect than correct responses. Therefore, these results do not need the formal application of a test of statistical significance to show that as a group these medical students could not tell when they were tasting light or bitter ale. In fact only 46% of responses (41 out of 90) were correct: we would want over 50% at least before making a claim for overall discriminatory abilities.1

Despite the overall negative answer to our original question, the results are worth more careful examination for clues that may lead to improved design in the future.

Is it easier to identify light ale or bitter ale correctly? The results could hardly be more similar: 47% correct

---

1 The expected distribution of the proportion of students with 0, 1, 2, . . . correct answers under the null hypothesis is given by the binomial expansion of $(\frac{1}{2} + \frac{1}{2})$. Similarly, the expected distribution of the number of students with five correct answers (in a series of experiments) under the null hypothesis is given by the binomial expansion of $(\frac{1}{2} + \frac{31}{32})$ Thus, in just over 10% of experiments two or more students would be expected to give five correct answers by chance, again indicating that our result in not significantly different from guessing alone.
answers when light ale was tasted against 44% for bitter.

Are there more, or fewer, correct answers as the experiment proceeds – that is, does learning play some part or do the taste perceptors become fatigued? Here the results are equivocal: there were 11 correct answers out of 18 on the first tasting followed by seven, eight, six, and nine. If anything, there is a slight suggestion that fatigue is more probable than learning. A longer interval - or at least a mouthwash, which was not provided - between tests might be of value.

There are other psychological aspects to such an experiment which may be important. First, only two students gave answers including four or more of one type of beer, whereas five students in fact received such an arrangement of beers. All the others, of course, gave answers including three of one and two of the other type, indicating an apparent expection of a three to two ratio of the beers. Secondly, only one student gave the same answer on the first two tastings, although in fact seven of the 18 were given the same beer on these occasions. This again indicates chance expectation affecting their judgement. The first of these points suggests that the design of the experiment might have been ‘improved’ by giving a prearranged number of each type of beer; for instance, three light and three bitter arranged randomly for each student. This restriction would, of course, affect the size of experiment necessary for a statistically significant result, since the number of possible answer permutations would be limited. Thus, answers with a four to two, five to one, or six to zero spilt would automatically be ruled out, and the chance of guessing the correct arrangement would therefore be increased. If this chance of being correct under the null hypothesis is not below the desired level of statistical significance, we would wish to enlarge the experiment.

Discussion

The who, what, and when of teaching statistics to medical students are vexed questions which the Todd report (Royal Commission on Medical Education, 1968) goes some way towards answering. As an applied subject it seems preferable that it should be taught in context by someone competent in both fields, but this is not always possible. From the statistician’s viewpoint the first year pre-clinical class does not offer so much of a challenge to his medical vocabulary; perhaps later the student’s clinical knowledge and involvement is somewhat of an obstacle.

Thus, the discussion of scientific logic and analysis could well be carried out on ground familiar to both sides, incorporating practical work to aid understanding. The relevance to medical experimentation should be outlined in some detail simultaneously. The field of beer drinking appeared to be suitable meeting ground; although, a priori, the students felt strongly that such an experiment as described was unnecessary because ‘Of course we can tell light from bitter!’ The results described, then, served well to highlight the importance of the concepts being taught, being a setback to morale for most of the participants and a strong incentive to understand more the logical arguments involved.

The relevance of the principles involved in this experiment, and other features of improved design such as stratification and double blindness, assumed a new significance to the group, whose initial approach is notorious not over-enthusiastic. The further discussion of drug and therapeutic trails, measurement and assessment of biological and observer vari-ability, causative factors in disease and other topics, became more lively and interested. Perhaps a blow to the ego is after all the best incentive to learning.

Summary

An experiment in beer tasting among 18 first year pre-clinical students showed that only two appeared to be able to discriminate light from bitter ale correctly. As a whole the group did worse than would be expected by guessing at random on each tasting trial. Some psychological biases, one possibly as a
result of the design, are indicated by the results and a possible remedy suggested.

The danger of acceptance of ‘views’ in clinical medicine without confirmatory experimental evidence was brought out by the students own failure to confirm a hypothesis (or ‘view’) that they believed beforehand to be beyond question.

Acknowledgements

I am grateful to Dr. C.R.B. Joyce of the Department of Pharmacology, The London Hospital Medical College, for criticizing an earlier draft of this paper. I wish also to thank Miss Pat Ward for ably assisting in carrying out the experiment, and the students who participated.

References