A similar path of reasoning can be used to conclude that if the median is to be used as the estimator of central location then the median absolute deviation from the median (M.A.D.) is a suitable measure of spread.

University College of Swansea

Reference

Hart, A. E. (1984). How should We Teach the Standard Deviation. *Teaching Statistics*, **6**, 24.

Lotteries and Probability: Three Case Reports

JAMES A. HANLEY

Keywords: Teaching; Laws of Probability; Birthday Problem; Location Measures; Coincidence; Binomial Distribution

Introduction

Teachers often find that one of their most difficult taks is teaching the use of first principles to calculate probabilities for non-standard situations. Textbook examples are often contrived (balls from urns, die tossing, . . .). Even when examples are topical (e.g. the World Series is well timed for North American courses that begin in September) they either have limited appeal or make unrealistic assumptions.

Over the past five years, lotteries have become quite common. They offer a useful addition to the teacher's repertoire of applications of probability. They also give one a chance to see how probabilities are dealt with and reported in the outside world. In this note we share three case studies from the lay press and discuss their implications for teaching probability.

Case I

The following report in the *Montreal Gazette* on Thursday, September 10, 1981:

> *SAME NUMBER 2-STATE WINNER*
>
> Boston (UPI)—Lottery officials say there is 1 chance in 100 million that the same four-digit lottery numbers would be drawn in Massachusetts and New Hampshire on the same night. That's just what happened Tuesday.
>
> The number 8092 came up, paying $5,842 in Massachusetts and $4,500 in New Hampshire.

> "There is a 1-in-10,000 chance of any four-digit number being chosen at any given time", Massachusetts Lottery Commission official David Ellis said.
>
> "But the odds of it happening with two states at any one time are just fantastic", he said.

This example shows how different the after the fact calculation of a probability is compared to the classroom exercise and how quickly people 'jump to coincidences'. Although this problem is conceptually equivalent to the textbook example of having the same number show when two dice are thrown simultaneously, the reaction is quite different. Lottery officials are drawn to the specific number 8092 and visualize the sample space as having just four possibilities, only one of which seems to describe the phenomenon that has just occurred:

Outcome	*Massachusetts*	*New Hampshire*
1	8092	8092
2	8092	not 8092
3	not 8092	8092
4	not 8092	not 8092

The pitfall is not so much in not knowing how to calculate a particular probability but in defining and calculating the wrong probability. Two viewing techniques can be used to help arrive at the correct one. The first, more traditional, one would ask the observer to imagine that neither drawing has yet taken place. One would lay out the grid of 10 000 by 10 000 possibilities and use the two elementary rules of multiplication and addition. The second would have the observer imagine that the Massachusetts draw had already taken place and the New Hampshire one is about to. Thus, the task would be seen as trying to match *whichever* number has already been drawn.

Case II

The second case involves the same Massachusetts Daily Lottery (the *Game*). On February 6, 1978, the *Boston Evening Globe* carried a feature on how bettors choose numbers. It was based on interviews with lottery official David Hughes and with a number of ticket vendors. As a side issue, it reported:

> During the Game's 22-month existence, the illegal numbers pool has switched its payoff from the race-track parimutuel pool to the legal number. In that period, no winning number has ever been repeated, although the same four digits have won a second time in different sequence. Hughes, the expert, doesn't expect to see duplicate winners until about half of the 10,000 possibilities have been exhausted.

Needless to say, these statements prompted a number of letters to the editor, pointing out that either "the number drawing is rigged so as to prevent repeat winners" or else we had witnessed a very unusual event, since "the chance of there being no repeat in the roughly 660 plays is only 22 billionths of a percent".

This example raises several points. First, it is identical in structure to the "birthday problem" (references 1 and 3) except that there are $N = 10\,000$ possibilities instead of 365 and $d = 660$ drawings instead of the customary 20 to 30. The high probability of a repeat is often not appreciated until the teacher (with a large enough class of say 35 to 40 to make it a safe bet) actually carries out the experiment. Only then is it apparent what the standard answer of 1 chance in 12 (for a class of 30) is based on: each student thinks of the probability as the chance that somebody will have the same birthday *as himself* rather than that *any two* students would share a common birthday. Indeed this narrowing of attention to 'just me' is very similar to the 'this specific number' trap in case study I.

Second, how does one calculate how many draws, d, it takes to produce a repeat? The median or other percentiles for d can be obtained either by brute force if the number N of possibilities is small, or by an approximation if N is larger (Feller, 1968). For example, if d is likely to be small compared to N, the median is close to $6/5\sqrt{N}$ (120 in our example). Formulae for the mean and mode of this distribution do not seem to have been published. The following argument can be used to derive the mode: Let

$$p_N(d) = [N/N][(N-1)/N]\ldots[(N-d+1)/N]$$

denote the probability that all d draws produce different numbers. Then the probability that the first duplicate occurs at the dth draw is the product of the probability that there were no duplicates in the first $d-1$ drawings and the probability that at the dth draw, one of the previous $d-1$ is duplicated. Denote this product $[(d-1)/N]p_N(d-1)$ by $f_N(d)$. It is at a maximum when

$$f_N(d) = f_N(d+1)$$

i.e.

$$[(d-1)/N]p_N(d-1) = [d/N]p_N(d)$$

i.e.

$$[(d-1)/N] = [(N-d+1)/N][d/N]$$

i.e.

$$d \simeq \sqrt{N}$$

or in our case

$$d \simeq \sqrt{10\,000} = 100$$

The derivation of the mean (close to 125 in our example) is left as an exercise for advanced students.

For the third point raised by this example, we return to our story to illustrate that the traditional reasoning either the model (sampling with replacement) is wrong (H_1), or we are looking at a rate event (H_0) is inadequate. In this case, the explanation was H_2: the data were incorrect! Apologetic Lottery officials announced one month later that there "had indeed been repeated numbers": seven separate numbers had repeated in the 22 months. "The misinformation was a sin of omission and a too-hasty glance at our own listing of previous winning number." The moral of the story is clear.

Case III

The final case concerns the Quebec Super Loto. Officials decided to pay out money accumulated from unclaimed prizes and at the same time help the ailing automobile industry by adding 500 cars as bonus prizes in the July 25, 1982 draw.

The *Montreal Gazette* reported on July 28:

> *$10 TICKET WINS BUYER TWO OLDS*
>
> Toronto (CP)—Antonio Gallardo has won two Oldsmobile Cutlass Supremes on a single $10 ticket.
>
> Gallardo, who had been shopping for a new car, was given the ticket by his sister, visiting from California.
>
> She bought him the ticket when she heard there were 500 cars being given as bonus prizes last Sunday.
>
> In Montreal, a Loto Quebec Corp. official said the chance of a single bonus number coming up twice is one in 46,181,926.

This variation on case study II illustrates a number of issues. First, the sampling without replacement traditionally used in raffles is not so easily accomplished with modern systems which draw each digit separately using numbered balls. Second, the viewpoint of the Loto Quebec official is interesting in that he sees the probability of 1 in 46 million from the standpoint of a customer. He used a Binomial calculation with $n = 500$ and $p = 1$ in 2·4 million (the number of tickets sold). If by a single bonus number he means any one customer (specified in advance), he is technically correct. However, from Loto Quebec's point of view, the chance that *some* number would be drawn twice is more like 1 in 20, still unlikely, but illustrating that rare events do indeed occur.

Third, both it and the Massachusetts lottery emphasize how clusters can be produced naturally by simple random fluctuation i.e. how lightning can strike twice. If one distributes $d = 500$ events over $N = 2{\cdot}4$ million households, one should not be too surprised to find some household (or group of households) receiving more than its fair share.

Discussion

The three case studies show how easily even people with some statistical training are led to erroneous reasoning, either because it supports the data (case study II) or because they would rather be confronted with the unexpected than the expected (cases I and III). All three involve an after the fact reasoning which subconsciously blocks out part of the relevant sample space. Given that most probability assessments in life are performed after the fact, it is clear that such assessments need to be handled very carefully.

McGill University, Montreal

References

1. Bissel, A. F. (1980). Breakdowns and Birthdays. *Teaching Statistics*, **2.1**, p. 15.
2. Feller, W. (1968). *An Introduction to Probability Theory*, Vol. 1, 3rd Edition. Wiley, New York.
3. Wilkie, D. (1981). Birthdays and Breakdowns Revisited. *Teaching Statistics*, **3.1**, p. 17.

Statistical Ideas in English Studies

N. C. SMEETON and EUNICE A. SMEETON

Keywords: Teaching; English; Opinion Poll; Road Accidents.

A major concern in the teaching of English Studies is the development of skills in the use of the English language, in order to equip the individual for the "outside world". Traditionally, this has been achieved by instruction in essay composition, letter writing and the comprehension of literature. However, with the changing demands of today's world, one should perhaps look at ways in which the English Studies lesson could be adapted, in order to increase its effectiveness.

The presentation of information through newspapers and television is one area which has undergone great change. Whereas at one time vague statements were often considered sufficient, arguments are now frequently backed up by figures. The opinion poll has become a prominent feature of political elections. Advertisers increasingly use surveys and statistical trials in order to promote their brands. It is becoming more necessary to be able to understand numerical data and interpret it into plain English.

Such interpretation requires the ability to understand the results, to question whether the method used to obtain them is valid and to draw suitable conclusions. Some training in these areas could be incorporated into the English Studies lesson. We shall look at how this could be done for the three areas separately.

The understanding of the results, as presented, is the most fundamental step. In tabular or graphical form they can be very confusing to those unfamiliar with such methods of presentation. As an example, consider annual road traffic accident figures for a small town, collected over several years and tabulated by type of accident. Training should be given to enable one to pick out such prominent features as the most frequent types of accident (by looking for the highest numbers in the table) and any major changes in the relative importance of different accident types over time. Initially, a step by step example should be given by the teacher. Along with this, the development of skills in English should always be at the forefront. Once the pupil has seen the main features from the information, well constructed sentences should be used in a written expression of the results.