The normal theory tolerance interval counterpart to Formulas (5), (6), and (7) is to use intervals based on endpoints of the form

$$\bar{y} \pm \tau s. \tag{16}$$

Odeh and Owen (1980) present rather extensive tables of constants τ useful for both one- and two-sided tolerance intervals for various fractions p of an underlying distribution at various confidence levels $\gamma \times 100\%$. Intervals (16) should be a serious part of introductions to statistical methods.

But beyond (16) is the possibility of providing at least one-sided tolerance-interval methods for essentially *any* instance of the constant variance normal linear model. It is a first-year graduate level exercise to show that under the conditions leading to (15), a $\gamma \times 100\%$ one-sided tolerance bound for a fraction p of all future observations at conditions x can be made using one of the endpoints

$$\hat{y}_x \pm \tau_x s, \tag{17}$$

where

$$\tau_x = A_x \cdot Q_{t(\delta_x,\nu)}(\gamma) \tag{18}$$

for $Q_{t(\delta,\nu)}(\cdot)$ the inverse noncentral t cdf for noncentrality parameter δ and degrees of freedom ν, where $\delta_x = Q_z(p)/A_x$. [$Q_z(\cdot)$ is standing for the inverse standard normal cdf.]

These days many statistical packages can be used to provide the noncentral t quantile needed in (18) [for example, the SAS Supplemental Library function TINV(P, DF, NCT) could be used]. But there is also an old route to an explicit approximation of the limits (17). That is, it is a relatively simple exercise to show that under the conditions leading to (15), approximation of the distribution of $\hat{y}_x + ks$ by a normal distribution with mean $E\hat{y}_x + k\sigma$ and variance $\sigma^2(A_x^2 + k^2/2\nu)$ leads to the conclusion that approximate $\gamma \times 100\%$ one-sided tolerance bounds for a fraction p of all future observations at conditions x can be made using (17) and

$$\tau_x \approx \frac{Q_z(p) + A_x Q_z(\gamma)\sqrt{1 + \frac{1}{2\nu}\left(\frac{Q_z^2(p)}{A_x^2} - Q_z^2(\gamma)\right)}}{1 - \frac{Q_z^2(\gamma)}{2\nu}}. \tag{19}$$

This kind of approximation is common (at least in one-sample contexts) in the statistical quality control literature and is traceable to Jennett and Welch (1939).

Formula (19) is usually adequate for practical purposes. Thus, even if one does not have access to software needed to use (18), it is possible to specialize (17) to each one of the spectrum of linear models used in introductory courses, and to thus provide reasonably explicit normal theory tolerance bounds.

7. SUMMARY

There is, of course, no end to the list of statistical interval methods that one might potentially recommend for inclusion in a first course. (I even believe *simultaneous* interval methods to have a place!) Most readers will, however, probably conclude that the suggestions here go beyond what they are presently willing to attempt. But it is hoped that this discussion will provoke *some* thought and movement on the part of a number of instructors toward a more comprehensive early teaching of "the other intervals."

REFERENCES

Hahn, G. J. (1970), "Statistical Intervals for a Normal Population," *Journal of Quality Technology*, 2, 115–125; 195–206.

Hahn, G. J., and Meeker, W. Q. (1991), *Statistical Intervals: A Guide for Practitioners*, New York: John Wiley.

Jennett, W. J., and Welch, B. L. (1939), "The Control of Proportion Defective as Judged by a Single Quality Characteristic Varying on a Continuous Scale," *Journal of the Royal Statistical Society*, Ser. B, 6, 80–88.

Odeh, R. E., and Owen, D. B. (1980), *Tables for Normal Tolerance Limits, Sampling Plans and Screening*, New York: Marcel Dekker.

Scheuer, E. M. (1990), "Let's Teach More About Prediction," in *Proceedings of the Statistical Education Section, American Statistical Association*, pp. 133–137.

Whitmore, G. A. (1986), "Prediction Limits for a Univariate Normal Observation," *The American Statistician*, 40, 141–143.

Jumping to Coincidences: Defying Odds in the Realm of the Preposterous

JAMES A. HANLEY*

1. INTRODUCTION

The calculation of probabilities is central to statistical inferences; however, teachers find it difficult to guide students, especially those who are not trained in probability, on how to set up correct probability calculations. Probabilities of seemingly rare events that are assessed *after the fact* are especially problematic; students employ selective vision and ignore other similar events in the sample space that would have prompted the same surprise and should therefore have been included in the calculated probability.

In a small publication aimed mainly at high school teachers, Hanley (1984) described three examples where probability specialists themselves have been "near-

*James A. Hanley is Associate Professor, Department of Epidemiology and Biostatistics, McGill University, Montreal PQ, H3A 1A2, Canada.

This work was supported by an operating grant from the Natural Sciences and Engineering Research Council of Canada. Josée Dupuis and Julie Bérubé provided helpful comments on the article.

sighted" in assessing or predicting usual and unusual events generated by state lotteries. In one example, a lottery official offered "data" which one *should expect* from a fair lottery; unfortunately, the logic used to "predict the usual" was faulty. In the two others, the unusual (what one *should not often expect*) was calculated to be much more unusual than it really was. These types of calculations are not new; two of the three cases were variations on the "birthday problem." Probability calculations such as these are taught in many college courses that touch on probability; however, they are couched in *prospective* (*before the fact*) terms and involve nameless coins, dice, beads, and urns.

Since then, a fourth unusual lottery event has been reported, this time from the New Jersey State Lottery. This story, and the official statistical reaction to it, were carried in worldwide publications. Again, the reported reaction was based on faulty probability calculations. These four faulty lottery calculations, and several new ones involving unusual births and birthdays, have prompted me to bring together in one place my "case series" and share it with a larger readership. Along with Paulos (1988), I argue that these variants of a common *probability blind spot* are not sufficiently appreciated and that we as teachers need to emphasize proper after-the-fact probability calculations a lot more; I relate my experiences in trying to do so to nonstatistical students.

A recent article by Diaconis and Mosteller (1989) dealt with a wider range of problems of coincidences and developed methods appropriate to students of probability. [This prompted a newspaper article by Kolata (1990), the science writer for the *New York Times*.] My intended readership here is teachers of wide-audience courses in statistics, since our aim should be to get students in all disciplines to be suspicious and ask good questions when confronted with the urge to calculate (or try to calculate) preposterous probabilities.

2. CASE STUDIES INVOLVING LOTTERIES

Unlike the birth examples to be discussed later, lotteries have the special advantage that their probability structure is usually well laid out, and few simplifying assumptions are needed. In all, four lottery examples are given. Since the first three have been discussed in detail by Hanley (1984), and space is limited, the commentary on them will be shorter than it might otherwise be; the most recent case, used as an example of the *law of truly large numbers* by Diaconis and Mosteller, is discussed in a little more detail.

Lottery Case 1 (from the *Montréal Gazette* on September 10, 1981)

Same Number 2-State Winner

Boston (UPI)—Lottery officials say that 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 drawn 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.

Although this problem is conceptually equivalent to the textbook example of throwing two dice and having them show the same number, the reaction to it is quite different: lottery officials are drawn only to the specific number 8092. Two techniques can be used to encourage students to "take off their blinders." Traditionally, the calculation is set up by laying out the grid of 10,000 by 10,000 possibilities and using the elementary rules of multiplication ($10^{-4} \times 10^{-4}$) and addition (summing this product over the 10^4 diagonal entries) to arrive at the correct answer of 10^{-4}. (In the case of the two dice, it is not clear to students whether the choice of first or second die is arbitrary—some textbooks make the dice different colors to make them more "real"; having two states makes the counting a lot easier.) The second approach is to imagine that the Massachusetts draw has already taken place and that the New Hampshire one is about to. Thus, the requirement for a newspaper story is that the New Hampshire match *whichever* number has already been drawn.

Of course, one could expand the calculation posed in the first paragraph of the story (just the specific states of Massachusetts and New Hampshire) to a much broader one by asking what are the chances of a match among *some two* of the *several* states with daily four-digit lotteries (the event implied by the title of the story)? The method of calculating this increased probability is a variation on that used in the following Lottery Case 2 (the number of states in Case 1 plays the same role as the number of draws in Case 2) and so will not be dealt with separately here.

The teacher can also use this expansion of Case 1 to distinguish between the probability of an event (a) *occurring* and (b) *occurring and being noticed and reported*. The news media in the adjacent states of Massachusetts and New Hampshire report the winning numbers in both their own and their next-state lotteries. This overlapping coverage, and the fact that residents of one state work in, and play the lottery in, the adjacent state make it more likely that a coincidence is noticed than, for example, if the match was between geographically distant Massachusetts and New Mexico or between states that are lexically distant in *USA Today's* listings, alphabetically by state, of the winning lottery winners. The example can help teachers to emphasize that one must take over- and under-reporting "filters" into account when, for example, judging the true incidence of a disease or other adverse event from spontaneously reported "cases."

Lottery Case 2 (from the *Boston Evening Globe* of February 6, 1978)

An article reported an interview with lottery official David Hughes on how bettors choose numbers in the above-mentioned Massachusetts Daily Lottery (the *Game*), which is played daily:

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 sequences. Hughes, the expert, doesn't expect to see duplicate winners until about half of the 10,000 possibilities have been exhausted.

In Case 1 (and as we will see in Case 4), players might be led to infer that unexpected events happen more frequently than expected! In Case 2, a lottery official reassured players that his lottery was fair (H_0) and offered "data," which were well within what *was* expected, to support that claim. A number of letters to the *Boston Globe* quickly pointed out that either (H_1): "the number drawing is rigged so as to prevent repeat winners" or else that we had witnessed a very *unusual* event, since (under H_0) "the chance of there being no repeat in roughly 660 plays is only 22 billionths of a percent." The reactions of these statisticians illustrate the dangers of believing in a two-hypothesis world. The explanation was H_2: Mr. Hughes's 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 numbers." I like to describe this as a Type III error!

Lottery Case 3 (from the *Montréal Gazette* of July 28, 1982)

Once or twice a year, the Quebec Super Loto pays out money accumulated from unclaimed prize-money by adding 500 cars as bonus prizes. Instead of mechanically drawing the large list of winning numbers from the 2.4 million tickets sold for each drawing, the Loto generated the 500 winning numbers using a computer, and published them—in the order drawn—in the local newspaper. After one such special drawing, the newspaper reported:

$10 Ticket Wins Buyer Two Olds

Toronto (CP)—Antonio Gallardo has won two Oldsmobile Cutlass Supremes on a single $10 ticket. A Loto Quebec Corp. official said that the chance of a single number coming up twice is one in 46,181,926.

Evidently the numbers were being drawn *with* replacement! The unsorted list of 500 numbers made it difficult for lottery officials to check for *any* duplicates, although Mr. Gallardo probably had no difficulty in finding that *his* number was there twice! In order to emphasize that "rare events do happen, and the small probability should not deter you from playing," it makes good marketing sense for lottery officials to calculate the small probabilities of such an event from an *individual customer's* viewpoint. The official correctly calculated "one in 46,181,926" as the binomial probability of Mr. Gallardo obtaining two successes when $n = 500$ and $\pi = 1$ in 2.4 million. However, in this instance, what was an unexpected bonus for Mr. Gallardo was a major embarrassment for the lottery corporation, which seemingly failed to appreciate that the chance of *some* number being selected twice is not negligible. The probability calculation is identical in structure to the birthday problem, but with $N = 2.4$ million rather than 365, and $n = 500$ rather than the customary 20 or 30. For the probability of no repeat, the 500-term product $\Pi(1 - i/N)$, with $i = 0$ to 499, can be very closely approximated by $(1 - 0.5n/N)^n$ to give .95, leaving the probability of a repeat at around .05. In other words, the lottery officials should have expected their "one in 46,181,926" event to occur on average once in 20 draws. In fact, it happened well before the 20th draw; since then, the lottery corporation draws 500 *distinct* winning numbers—and publishes them in numerical order!

Lottery Cases 2 and 3 are examples of the *duplicate birthday problem*. Of the many explanations of how our intuition lets us down in this problem, I prefer the one attributed to Cornfield (Slonim 1960). He said, in effect, that the average person sees the problem as "What are the odds that any of the other $n - 1$ has the same birthday as *mine*?", whereas he more properly should ask, "What are the odds that any one of the n has the same birthday as *any other one of the n*?" However, in a classroom, having students call out their own birthdays makes the task of looking for a duplicate too easy, since each student simply continues to check his or her own against those being called out. The lottery corporation's task of checking through 500 (Quebec) or 660 (Massachusetts) nameless numbers in an unsorted list is a lot more boring, and takes one person much longer to do; however, the amount of work involved is a clue to the magnitude of the probability. To illustrate the high probability of duplicate birthdays, I use a spreadsheet to generate several columns of 23 random birthdays, and ask students to check for duplicates within each column and to explain how they cover all possibilities. They quickly see that checking the first entry against entries 2 to 23, then the second against 3 to 23, and so on involves a lot of work, and I explain that every possibility they check adds to the probability, that is, the numerator is far larger than they had originally thought.

Lottery Case 4 (from the *New York Times* of February 14, 1986)

Odds-Defying Jersey Woman Hits Lottery Jackpot 2d Time

Defying odds in the realm of the preposterous—1 in 17 trillion—a woman who won $3.9 million in the New Jersey state lottery last October has hit the jackpot again and yesterday laid claim to an additional $1.5 million prize . . .

She was the first two-time million-dollar winner in the history of New Jersey's lottery, state officials said. They added that they had never before heard of a person winning two million-dollar prizes in any of the nation's 22 state lotteries.

For aficionados of miraculous odds, the numbers were mind-boggling: In winning her first prize last Oct. 24, Mrs. Adams was up against odds of 1 in 3.2 million. The odds of winning last Monday, when numbers were drawn in a somewhat modified game, were 1 in 5.2 million.

And after due consultation with a professor of statistics at Rutgers University, lottery officials concluded that the odds of winning the top lottery prize twice in a lifetime were 1 in about 17.3 trillion—that is, 17,300,000,000,000.

It is interesting how officials use their statistical consultants to "jump to coincidences," that is, how correct probabilities in the penultimate paragraph are "tele-

scoped" to "odds of winning twice in a lifetime." As Samuels and McCabe (1986) argued, the *type* of event that occurred was far from the miraculous; in fact, it was to be expected.

The final calculation ignored several pertinent facts [Hanley, J.A. "Jumping to coincidence" (unpublished) letter to *New York Times*, February 24, 1986] first in calculating the odds in Mrs. Adams's case, and second in estimating how soon *someone, somewhere* (not necessarily Mrs. Adams, or even in New Jersey) would hit the jackpot twice.

First, in Mrs. Adams's case, the calculation assumed that she played only two weeks—once when the game was "6/39" with 3.2×10^6 possibilities, and once when the game was "6/52" with 5.2×10^6 possibilities—and that she bought just one ticket for each of these two weeks, making the odds 1 in $3.2 \times 5.2 \times 10^{12}$, or 1 in 17.3 trillion. ("k/n" refers to a game where a player chooses k numbers; the biggest prize is shared among those players whose k choices match exactly the numbers on k balls drawn without replacement from n balls numbered 1 to n; the probability of getting x correct out of k is given by the hypergeometric distribution with parameters k and $n - k$.) In fact, she had bought several tickets each week for some years and had raised her purchases after she won the first time. Each week, with about 5 million possibilities, a player who plays just five tickets a week (or three tickets in the older, "6/39," version of the game) has a probability of about 1 in a million of winning the big prize. In four years or about 200 games, the same type of binomial (or Poisson) calculation as in Lottery Case 3, but with $n = 200$ and $\pi = 10^{-6}$, puts the approximate chance of not winning any game at $\exp(-n\pi)$ or practically 100%, of winning once at $\exp(-n\pi) \times n\pi$ or about 1 in 5,000 (i.e., 200 in a million), and of winning twice at $\exp(-n\pi) \times [n\pi]^2/2$ or about 1 in 50 million. If one plays for a "lifetime" of just under 30 years ($n = 1{,}500$), the odds of a "double" improve to 1 in a million.

Second, odds of 1 in 50 million, or even 1 in a million, are still daunting for any one player. However, if a million persons in New Jersey play five tickets a week for a lifetime, then some one of them can be expected to hit the jackpot twice. As the saying goes, the double "has to happen to *some*body, *some*time" and after all, "1 in a million" means exactly that—one two-time winner in one million lifetimes. Similarly, if 50 million people nationwide play five tickets weekly, one of them can be expected to win twice in just four years.

The *Times* correspondent correctly put the lifetime odds given by the officials "in the realm of the preposterous." Such odds imply that one two-time winner can be expected only when every person in the entire world population had played the New Jersey lottery for 4,000 lifetimes! The fact that the game-playing population of New Jersey has already achieved this feat in a few short years should prompt one to reexamine the basis for the probability calculation. As was emphasized in my discussion of Case 2, obtaining a very small probability for an event that has occurred should prompt one to be suspicious of both one's data and of one's calculations.

If one knew how many, and how regularly, people play the nation's lotteries, how many tickets they buy, which numbers they choose more often, and whether they pool their tickets, one could calculate the real odds more accurately than those given here. Even with these refinements, however, the correct odds must remain in the "human" realm. Lottery officials and statisticians, either because such events are not commonplace, or because they like to calculate numbers with many zeroes, or because they think the public will buy more tickets, continue to make the odds of "interesting" events much more preposterous than they really are.

3. CASE STUDIES INVOLVING BIRTHS AND BIRTHDAYS

We report three examples. The first and third beg us to multiply small probabilities; the second one does the calculation for us.

Births Case 1 (from *USA Today* on March 4, 1987)

Two Sets of Quints, Same Day

It was a statistician's dream: two sets of quintuplets born the same day [Monday March 2, 1987] in the USA [one set of four girls and a boy, in Peoria Ill. and a set of five girls in Las Vegas]. But the experts could not agree on what statistic to use. Figures of one in 41 million, one in 70 million, and one in 85 million were tossed out Tuesday—and that was just for the birth of one set of quints. [The mother of the Nevada quintuplets, who were born 11 weeks early, didn't use fertility drugs; the mother of the Illinois quints had taken the fertility drug Pergonal].

The use of fertility drugs certainly complicates an already complex, and ill-defined, event. We leave it to the reader to try to define what are the relevant probabilities to calculate. However, we believe that they must include specification of the time-horizon (1987? a span of 10 years? 50 years?) and of which day (presumably any day, not just March 2, would have produced the same headline) along with a much larger probability of quints with the use of fertility drugs.

Births Case 2 (from *National Enquirer* on June 28, 1990)

4 Sisters Beat 1 in 17 Billion Odds—They All Share the Same Birthday

August 3 is a grandslam event for Mary Wohlford—her first four daughters were born on that date in four different years. The odds of that happening are a staggering 1 in 17 billion. Her first August 3 child, Connie, arrived in 1949. She was followed by Sandra in 1951, Ann in 1952 and Susan in 1954.

All were born in Freeport, Ill. and delivered by the same doctor in the same hospital in the very same room. "The doctors and nurses were amazed, but it was not planned that way, and the girls weren't all due August 3" said the mother.

But the August 3 streak ended after the parents moved their growing family from Freeport to Dyersville Ill. "Maybe there was something in the Freeport water" jokes the mother. "After we moved in 1955, we had four more girls over the next nine years and none of them were born on August 3rd."

This example was brought to me by Baird Smith, a physician in my statistics course in our 1990 summer

school (as part of the course assignment, each student was required to construct an exercise, from real data, which could be used in next year's course). Even without insisting on the "same doctor, same hospital, same room," the story allowed him to devise a large variety of calculations and restrictions: four *consecutive* sisters with same birthday, *any day*; four *consecutive* sibs with same birthday, *specified day*; *any* four sisters with same birthday; and so on which showed that the *National Enquirer* took considerable liberties with its calculation of $365^4 = 17$ billion. In the usual birthday problems, the fact that birthdays are not quite uniformly distributed throughout the year is a minor problem. Here, however, as in the previous case study, the assumptions are more critical, if somewhat difficult to quantify, particularly in relation to the time window available for a birth in 1952, and in relation to the parents' possible deliberate efforts to keep the streak going [for example, the probability that four children, all due on a certain date, would all arrive on the same date (not necessarily the due date) is nontrivial, and is bigger still if gestational ages are positively correlated within the same mother]. In any event, when one sees a small probability such as 1 in 17 billion, shouldn't one be suspicious and stop to think "how many families with eight (or even four) children have there been in the world since people began to use calendars?" I estimate it is probably of the order of a few billion—using the size of the current world population, assuming a generation length of 25 years, and assuming from the geometric rise in the world population that most of the world's lifetime population was born in the 20th century.

Births Case 3 (from the *Montréal Gazette* in May 1989)

Double Trouble in Moose Jaw School
(caption to a photograph showing six sets of twins)

Every morning, teachers at Prince Arthur school in Moose Jaw, Saskatchewan see double—and it's not because of what they were up to the night before. Six pairs of identical twins attend the school, which has an enrollment of 375. Identical births occur once in 270 births.

I use this example in class to illustrate how one can visualize the Poisson distribution using the very useful cell occupancy approach. I ask students to think of randomly assigning the approximately 10,000 twins in 2,700,000 births in Canada in a space of five to six years to 7,200 schools of size 375 each, and to imagine how many schools will receive 0, 1, 2, . . . Students quickly agree that if there are to be some 0's, then there must be some 2's and 3's, and a few even bigger clusters, in order to have an average of $10{,}000/7{,}200 = 1.39$ twin pairs per school. (The Poisson distribution should also apply in the United States; the numbers of schools and twin pairs would be 10 times bigger, but the mean per school would remain the same.) I use a microcomputer to simulate this distribution in real time, so that students see the twin pairs "piling up" one by one in the various schools. Because I cannot represent all 7,200 schools on the computer screen, I scale down the problem by a factor of 20, and use a grid of $7200/20 = 360$ boxes or cells to represent schools. At any stage of the assignment process, the numbers in each of the cells represent the number of times that each cell has been "visited." For each of the $N = 10{,}000/20 = 500$ "visits" (each one representing a twin pair), the target cell is chosen randomly (with replacement) from the numbers 1 to 360, and that cell's occupancy is updated (incidentally, when the run is completed, we are left with a table of random numbers with a Poisson distribution). Finally, in order to stimulate discussion of the enormous difference between using *named* and *unnamed* towns, I ask students if the headline would be any less remarkable if it read "Double trouble in 'Anytown Canada' school" and what the implications of a "write the headline first, fill in the name after the fact" policy would be. This example is particularly helpful in teaching students how to think about random disease clustering in epidemiology (imagine "twin pairs" changed to "childhood cancers").

This general-purpose cell occupancy program can also be used, with selectable numbers of cells and numbers of visits, to simulate duplicate birthdays, and to show such phenomena as the "multiple events" discussed in section 7.1.3 of Diaconis and Mosteller. It is programmed in QuickBASIC™ and is available from the author.

4. CONCLUDING REMARKS

Teaching probability continues to be a difficult challenge. These recent examples reinforce the need to be wary of coincidences and to be sure that we do not limit our field of vision in counting the types of events that evoke surprise. The increased reporting of lottery and other human-interest events, and the availability of "live" computing in the classroom are two new ways to make the subtleties of coincidences, and the dangers in naive "after the fact" calculations, easier to understand. By adding these seven case studies to each teacher's repertoire, by collecting and using the many more examples that occur locally, and by use of spreadsheet and other more special-purpose software to produce live simulations, I hope that we can teach probability better and that our students will be less likely to unquestioningly accept inhuman or preposterous probabilities.

Postscript (July 2, 1990) On Friday June 29, 1990 Dave Stewart of the Oakland Athletics and Fernando Valenzuela of the Los Angeles Dodgers pitched no-hitters. It was the first time in major-league history that no-hitters were pitched in each league on the same day. Can we reasonably expect to see a repeat in our lifetimes? Imagine a 180×2 grid representing the two leagues and the approximately 180 playing days in a year; over this one-year grid one randomly distributes say four no-hitter games. The chance that all of them will fall on different days is $(358/360)(356/360)(354/360)$. Either one of the remaining possible duplicates, namely "same day, same league" or "same day, different leagues," although they would be expected to happen

as frequently as once in 30 years (more frequently if no-hitters are not distributed with uniform intensity over the season) would surely become a headline!

[*Received June 1990. Revised August 1991.*]

REFERENCES

Diaconis, P., and Mosteller, F. (1989), "Methods for Studying Coincidences," *Journal of the American Statistical Association*, 408, 853–861.

Hanley, J. A. (1984), "Lotteries and Probabilities: Three Case Reports," *Teaching Statistics*, 6, 88–92.

Kolata, G. (1990), "1-in-Trillion Coincidence, You Say? Not Really, Experts Find," *New York Times*, Feb. 27, C1.

Paulos, J. A. (1988), *Innumeracy: Mathematical Illiteracy and its Consequences*, New York: Hill and Wang.

Samuels, S. M., and McCabe, G. P. (1986), "More Lottery Repeaters Are on the Way," *New York Times*, Feb. 27, A22.

Slonim, M. J. (1960), *Guide to Sampling*, London: Pan Books Ltd., p. 25.

Using Lottery Games to Illustrate Statistical Concepts and Abuses

RICHARD A. PAULSON*

Two applications of the analysis of lottery games as an instructional tool are described. First, the use of lottery examples in the statistical areas of probability and decision analysis are discussed. Second, two different systems for predicting lottery numbers are examined, along with a discussion of the statistical concepts that are being inappropriately utilized in each case. An analysis of these misuses of statistics can be an interesting vehicle for explaining statistical ideas to students in introductory classes.

KEY WORDS: Lottery games; Randomness; Statistical misuse.

Over the past decade lotteries have become a popular source of revenue generation for state governments. The increase in the number of states offering lotteries makes the study of these games an area of fairly general interest to a variety of students. There are a number of statistical areas for which the analysis of lottery activities is relevant. This article discusses some of those areas. It particularly focuses on systems for predicting future lottery numbers, and how an examination of such systems can be useful in explaining the abuse of statistical procedures to students.

Most lotto games involve the random selection of six numbers from a pool of somewhere between 39 and 54 numbers. These differ from instant lottery games where the payoffs are typically smaller and immediate. For these lotto games the order in which the numbers are drawn is not important and the numbers on different tickets are not necessarily distinct. If the six numbers on a purchased lottery ticket match the six selected numbers, then the holder of that ticket wins (or shares in winning) the first prize for that lottery game. These "jackpots" can be worth millions of dollars, thus creating an incentive for people to participate in this gamble. Monetary prizes are also awarded for picking less than six numbers correctly. Usually, four or five correct selections will return money to the ticket holder. Payoffs are on a parimutuel basis with a certain percentage of the total revenue pool being appropriately distributed to the holders of winning tickets. Typically, state-run lotteries return approximately 50% of all money collected to the ticket purchasers in the form of prizes. The other 50% is kept for expenses and profit. Shenkin and Wieschenberg (1985) explain that in real terms the percentage return to the ticket purchasers may be lower since most jackpots are paid out over a series of years (thus decreasing the current value of the winnings) and since winnings are subject to income taxes (while losses are not deductible from ordinary income).

Lottery games can have a variety of interesting academic uses. What is the probability of selecting the correct six numbers from a pool of 54 numbers? The number of different ways of selecting r objects from n objects, ignoring the order of selection, is

$$C(n, r) = \frac{n!}{r!\,(n - r)!}.$$

Thus, there are $C(54, 6) = 25{,}827{,}165$ different combinations of six numbers taken from a total of 54. Only one of these combinations will match the six selected numbers. So, the appropriate probability is 1/25,827,165, or .0000000387. What is the probability of selecting five of the six numbers that are drawn? There are $C(6, 5) =$ six ways of choosing five winning numbers out of six and $54 - 6 = 48$ ways of choosing one nonwinning number. Thus, there are $6*48 = 288$ combinations which yield five of the six correct numbers for a probability of 288/25,827,165, or .00001115. What is the probability of selecting four of the six numbers that are drawn? There are $C(6, 4)*C(48, 2) = 15 \times 1128 = 16{,}920$ combinations that yield four of the six correct numbers for a probability of 16,920/25,827,165 or .000655.

A lottery example with expected return computations is illustrated in Chernoff (1981). He describes the application of some basic probability concepts in analyzing (and attempting to beat) the Massachusetts lotto game.

*Richard A. Paulson is Associate Professor, Department of Business Computer Information Systems, College of Business, St. Cloud State University, St. Cloud, MN 56301-4498.

