METAMETRICS, INC.
Bringing Meaning to Measurement

Review of Data Analysis of Insider Ontario Lottery Wins

By Donald S. Burdick

Background

A data analysis performed by Dr. Jeffery S. Rosenthal raised the issue of whether retail sellers of tickets in the Ontario lottery were winning major prizes at an excessively high rate. If so, this issue is a matter of serious concern to the extent that the integrity of the process for determining major prize winners is called into question. Dr. Rosenthal's findings were contested in a study conducted independently at the behest of the Ontario Lottery and Gaming Corporation. This document is a critical review of these two studies.

Framing the Issue

When an individual purchases a ticket in a lottery, the presumption of fairness implies that the probability of winning a prize should not depend on who the individual is. In particular, the chance of winning a major prize, i.e. $50,000 or more, should not depend on whether or not the individual is an insider. Given a data set containing information about prizes won compared to money spent on the lottery by both insiders and outsiders, a statistical analysis of that data may be conducted to see if the results are consistent with what the presumption of fairness implies. Such an analysis typically involves data, one or more statistical models, and a statistical analysis leading to uncertain inferences expressed in probabilistic terms. The case at hand is no exception.

The basic approach is as follows. The assumption that each dollar spent buys the same chance at winning a major prize implies that the ratio of major prizes won by insiders to those won by outsiders should on average be equal to the ratio of money spent by insiders to money spent by outsiders. The use of the phrase "on average" serves notice that this relationship is not guaranteed to be exact. In fact the opposite is true. The chance mechanisms built into the normal operation of the lottery virtually guarantee that the two ratios will not be exactly equal. The statistical analysis is designed to assess whether the discrepancy between the two ratios is beyond the limits reasonable expectation.

The basic procedure for accomplishing this assessment can be described briefly. The four quantities required to calculate the two ratios are the total expenditures, the expenditures by insiders, the total number of major prizes won, and the number of major prizes won by insiders. Equality of the two ratios implies that the number of major wins by insiders could be calculated by multiplying the ratio of insider expenditures to total expenditures by the total number of major prizes.

We call this product the expectation and note that normal chance variation will produce a discrepancy between the actual number of insider wins and the calculated expectation. The amount of the discrepancy that normal chance variation is likely to produce can be calculated using a statistical model called the Poisson model. In particular, the probability that the discrepancy will exceed any specified amount as a result of normal chance variation can be calculated from this model. Of particular interest is the result of this probability calculation when the specified amount is the discrepancy actually observed. The smaller this probability turns out to be, the harder it is to believe the observed discrepancy is the result of normal chance variation.

There are important questions to keep in mind when performing a critical evaluation of an inferential statistical analyses such as this, i.e.

1. How reliable is the data on which the analysis is based?
2. Are the statistical models appropriate in the case at hand?
3. Is the methodology employed in the analysis appropriate?
4. Are the inferences drawn from the analysis justified in the current context?

Questions 2 and 3 could be asked in connection with the Poisson model and its use as described above. In the current context these questions, in regard to the Poisson model and its use, can be confidently answered in the affirmative. Later we will encounter other portions of the analysis where these questions will resurface in connection with other statistical models. For now, the focus will be on question 1 because of serious issues in connection with the data used as input for the statistical analysis based on the Poisson model.

The Data Input

The issue to be addressed involves the comparison of two ratios. Four numbers are needed in order to calculate these two ratios. These four numbers are the total expenditures, the expenditures attributable to insiders, the total number of major prizes won, and the number of prizes won by insiders. The amount $2.22 Billion is the figure used in the Rosenthal analysis for total expenditures. It represents a yearly average over the 1999-2005 period. It includes expenditures by both insiders and outsiders. The Rosenthal analysis uses 5713 as the total number of major prizes won during the 1999-2005 period. Both of these numbers are presumed to be highly accurate. To complete the input for the analysis, values are needed for expenditures by insiders and for major wins by insiders. Both of these numbers are subject to uncertainty, which means the question of data reliability cannot be easily dismissed.

The issues arising from uncertainty about the expenditures by insiders have a much greater impact than those arising from uncertainty about the number of

prizes the insiders won. It is worthwhile spending some time though to investigate the source of the uncertainty about the number used in the Rosenthal analysis for prizes won by insiders. The number of major prizes won by all insiders in the 1999-2005 period is given as 214, and this number is presumably accurate. The Rosenthal analysis is focused not on all insiders but on the subcategory of insiders consisting of employees and owners of retail outlets that sell lottery tickets. The number of insiders in that subcategory is substantial, leading Rosenthal to estimate that 200 of the 214 major prizes were won by owners/employees of retail outlets. The statistical methods on which that estimate of 200 was based imply that there is some uncertainty associated with that figure. For the sake of simplicity, henceforth, the term "insider" will refer to the subcategory of insiders consisting of owners and employees of retail outlets.

Evidence of uncertainty concerning the expenditures by insiders and the substantial effect it can have is reflected in the Rosenthal report when it, in effect, uses six different values for that quantity. For convenience I'll designate these six quantities as 1, 1a, 2, 2a, 3, and 3a. Each of the six estimates of expenditures by insiders is obtained by multiplying an estimate of the total number of insiders by an estimate of the average amount expended per insider. Each of the two factors is subject to uncertainty. The different numbers correspond to differing estimates of the total number of insiders. The presence or absence of the suffix "a" indicates the presence or absence of an adjustment factor, which is also subject to uncertainty. The purpose of the adjustment factor will be explained shortly.

Estimate #1 of total expenditures by insiders is $13,338,500 obtained as the product of 36,050 by $370, where the first factor is the estimate of the total number of insiders and the second factor is the estimate of the average annual expenditure for insiders. The source of both of these estimates was data obtained from 200 retail locations in a random survey conducted by Fifth Estate. Both estimates are subject to uncertainty.

Estimate #1a of total expenditures by insiders is $23,609,145, which is the result of multiplying estimate #1 by an adjustment factor of 1.77. The motivation for the adjustment factor is concern that the per insider average annual expenditure figure of $370 may be too low because of underreporting. The factor 1.77 was obtained as an estimate from a "small additional survey" conducted by Fifth Estate. As such, it too is subject to uncertainty.

Estimate #2 of total expenditures by insiders is $22,200,000, which is the result of using 60,000 instead of 36,050 as the estimated total number of insiders. The number 60,000 came from court testimony and is unsupported by any other reference to a specific data source. Estimate #2a is $39,294,000, the result of multiplying Estimate #2 by 1.77.

Estimate #3 of total expenditures by insiders is $37,434,380, which is the result of using 101,174 instead of 36,050 or 60,000 as the estimated total number of insiders. The number 101,174 came from an exhaustive list of 10,911 retail locations which were classified into 12 categories called channels. For each channel the average number of insiders per location was estimated from a survey of "representative" locations. Estimate #3a is $66,258,852.60, the result of multiplying Estimate #3 by 1.77.

How Many Insiders?

Estimate #3 (or 3a) is over 280% of Estimate #1 (or 1a). The difference between these estimates of expenditures is the result of differing estimates of the total number of insiders. The difference is much too large to be dismissed, so a critical examination is in order of the way in which these estimates were obtained. In both cases the estimate of the total number of insiders is obtained as a product of the number of retail locations by an average number of insiders per location. Although both approaches have this basic feature in common, the methodology for implementing it is quite different. We will examine each with attention to the sources of uncertainty in the numbers used.

Estimate #1 takes a global approach using 10,300 as the total number of retail locations. This number is somewhat different from the figure 10,911 used in the process of obtaining Estimate #3, but this difference is understandable and probably inconsequential. It is likely that both figures come from complete records rather than samples. The first figure is reported as an average over multiple years and the second is most likely specific to a particular year, most likely 2006. The uncertainty associated with these numbers is minimal. It is possible that both are exactly right.

Uncertainty plays a major role, however, in the number of insiders per location. Rosenthal reports an average of 3.2 employees per location in a "random survey of 200 locations conducted by Fifth Estate", but gives no further details about the survey's methodology. In particular the following questions were not addressed in the Rosenthal report.

1. What were the sampling units and the sampling frame used in the survey?
2. What randomization procedure was used to select sampling units from the sampling frame?
3. What methods were used to elicit information from the selected sampling units?

These questions are important, but they involve technical matters and should be explained further for a lay audience. The phrase "random survey of 200 locations" suggests that the sampling unit was the retail location, not the

individual insider. If so, the sampling frame would consist in effect of a list of retail locations from which a random sample of 200 locations could be drawn. A questionnaire might then be used to elicit the information about the number of insiders at each of these 200 locations, but if so, what questions were asked and of whom?

Now, let's examine the basis for Estimate #3. It is based on 10,911 retail locations classified into 12 categories or channels. Rather than an estimate of an overall average number of insiders per location, an average per location was obtained for each channel which could then be multiplied by the number of locations in the channel to a obtain channel-specific total. The overall total number of insiders is then obtained by summing the twelve channel-specific totals. The channel-specific average number of insiders was obtained from surveys of locations "most representative of their channels", i.e. not from randomly selected samples. Although the use of a subjectively determined representative sample rather than a random sample does not necessarily yield a less accurate estimate, it can and often does lead to biased estimates.

Comparing the details of the two approaches brings the importance of the questions about methodology of the Fifth Estate survey into sharp focus. The channel-specific averages from the second approach range from a low of 4 for independent convenience stores to a high of 40 for supermarkets. None are as low as 3.2, the average of the 200 locations in the Fifth Estate survey. Were there any supermarkets among those 200 locations? The 731 supermarkets in the 10,911 locations are 6.7% of the total. If the sampling frame for the Fifth Estate survey included 6.7% supermarkets, one would expect to see about 13 supermarkets among the 200. Perhaps 40 is an overestimate of the average number of insiders at supermarkets, but presumably there are at least a fair number of supermarkets with 40 or more insiders working there. Were any of the 200 insider counts in the survey as large as 40? A glance at the data would easily answer this last question, but the 200 counts are not given in the Rosenthal report. As it happens we can answer that question anyway. Rosenthal reports the standard deviation of the 200 counts to be 1.65. It is a mathematical impossibility for any of 200 numbers which have an average of 3.2 and a standard deviation of 1.65 to be as large as 40.

In summary, the difference between Estimate #1 and Estimate #3 is the consequence of a large difference in the respective estimates of the total number of insiders. This large difference cannot be easily explained as the consequence of either the randomness of the Fifth Estate survey or distortions arising from the representative sample approach employed to reach Estimate #3, provided the sampling frame for the survey closely matched the 10,911 locations used for Estimate #3. If, on the other hand, the sampling frame were limited to convenience stores, consistent inferences about channel-specific insider totals

could be made from the two data sets, but the estimate of the overall total used in calculating Estimate #1 would be too low by a substantial margin.

Before turning to the uncertainties associated with the estimates of expenditure per insider, I should note that 3.5, not 3.2, was the number used for the average insider count per location in the calculation of Estimate #1. This increase from 3.2 to 3.5 produces an upward bias. It was done to reduce the chance that the value 3.2 calculated from the sample of 200 would prove to be an underestimate if the survey were extended to the entire sampling frame. However, if the sampling frame was limited to convenience stores, the downward bias resulting from that limitation would most likely overwhelm the upward bias produced by the increase. The Rosenthal report refers to the "sample of 200 convenience store owners/employees", which suggests that the sampling frame was indeed so limited.

How Much Does An Insider Spend on the Lottery

The six estimates in the Rosenthal report of annual expenditures by insiders are each obtained as the product of an estimated total number of insiders and an estimated annual expenditure per insider. Having discussed the uncertainties associated with the first factor, I turn next to the uncertainties associated with the second.

For each of the six estimates, the estimate of average expenditure is either $370 or $370 multiplied by 1.77. The uncertainties associated with each of these numbers will be examined, beginning with $370.

According to Rosenthal, the amount $370 was based on data from 200 insiders interviewed in the Fifth Estate survey. These insiders were asked how much spent they spent playing the lottery. The 200 answers formed the data base from which the estimate $370 was calculated. Since the Fifth Estate survey included 200 locations, it seems reasonable to infer that, although many locations had more than one insider, only one insider was interviewed at each location. That raises some methodological questions.

1. Was the insider to be interviewed selected at random from a frame listing all insiders at the location or was some other selection method used?

2. If the selection was random, what randomization device was used?

3. How were the questions about expenditures phrased?

There is an issue worthy of mention here, although I judge it unlikely to have a major impact in this case. When estimating total expenditures by insiders via a random survey, the most natural sampling frame to contemplate is a listing of all insiders. It has the property that every insider has the same chance of being included in the sample, which assures that the sample average will be an unbiased estimate of the population mean. If instead, locations are sampled at random and an insider is sampled at random from the selected location, not every insider has the same chance of inclusion in the sample. The insiders at locations with few insiders would be more likely to be in the sample than would insiders at locations with many insiders.

Moving on, we turn to issues which are likely to have more of an impact on the estimate of average expenditure per insider. Of particular importance is the issue of underreporting. The dollar figure $370 used in the calculations is based on self-reported expenditures from the Fifth Estate survey. This figure might well be too low because of underreporting. This possibility was recognized and addressed in the Rosenthal report. The means for addressing was a "small additional survey" of the general population conducted by Fifth Estate. The small survey yielded an average of $141.03 for self-reported annual expenditures. This value is below $249.44, which is based on the ratio of actual receipts to adult population and may be regarded as a reliable estimate of average annual expenditures for the general population. The ratio of 249.44 to 141.03 is 1.77, which is used as an adjustment factor in obtaining Estimates #1a, #2a, #3a in lieu of the corresponding estimates which use an unadjusted value of $370 for average annual expenditures by insiders.

The factor 1.77 comes from data in the small survey and, like the main survey, it is subject to uncertainty and questions about methodology. However, it clearly confirms the expectation that self-reported expenditures are likely to be low. Consequently, whatever the estimate of the total number of insiders, the value $370 is highly likely to be an underestimate of average expenditures per insider, which argues against the use of Estimates #1, #2, #3 in favor of the corresponding Estimates #1a, #2a, #3a.

The Rosenthal report did not address any issues arising from uncertainty in the adjustment factor 1.77. Moreover, the report contained less detailed information for the small survey than for the main survey. For example, Rosenthal reported an average expenditure of $476.31 for insiders in the main survey along with the number of self-reported figures from which that average was calculated and the standard deviation of those figures. For the small survey we have only the average value of 141.03.

We can make some speculative guesses about the missing information from the small survey as a means of getting a rough idea of the possible impact that random error in the small survey could have on the adjustment factor. For the

main survey the average of the self-reported expenditures was 476.31 with a standard deviation of 602.5. The ratio of 602.5 to 476.31 is 1.265. If that same ratio applied to the small survey data, the standard deviation of the numbers used to calculate the average of 141.03 would be 178.4. The number of respondents in the small survey is presumably less than in the main survey, so I'll guess that number to be 50. Dividing 178.4 by the square root of 50 yields a standard error of 25.23. It is quite possible for an estimate to be one or more standard errors too high. If we subtract the hypothesized standard error from 141.03, we get 115.8. The adjustment factor when 141.03 is replaced by 115.8 is 2.154 instead of 1.77. Rosenthal reports the expected number of wins by insiders derived from Estimate #3a to be 170.5=1.77*96.33. If we replace 1.77 by 2.154 in that calculation, we get 207.5 as a plausible expected number of wins by insiders, which is greater than the actual number.

Other Sources of Uncertainty

This review has addressed in depth some, but not all, of the sources of uncertainty in the numbers used in the calculations performed by Rosenthal. To address these other sources in depth would add substantially to the length of this review and be tantamount to overkill, but a few of these sources are worth at least a brief mention.

The adjustment factor for underreporting might not be constant at all expenditure levels. Someone who plays the lottery at a high rate may be more inclined to underreport than someone who plays less. Both the Fifth Estate survey and the one conducted at the behest of the Ontario Lottery and Gaming Corporation found the expenditure rate by insiders to be higher than the rate for the general population. If higher expenditure rates are associated with greater underreporting factors, then a small random telephone survey of the general population would yield a biased underestimator of the underreporting factor for insiders. If the underreporting factor is too low, the estimate of insider expenditures will be too low, also.

Some of the 200 insiders in the Fifth Estate survey admitted playing the lottery, but refused to say how much. Treating these nonrespondents as missing at random leads to negative bias in the average reported expenditures of those who did respond. Adjusting the average upward by one standard error may not be enough to compensate for this negative bias.

There are a number of ways to play the lottery, and they don't all have the same chance of winning a major prize. Perhaps insiders are better informed and play the games with better chances more often than the general public.

Group play could be having an effect. If ten persons pool their resources for a group lottery play and one of the ten is employed at a retail outlet, that one person

may be asked to make the purchase as a matter of convenience. If that purchase results in the win of a major prize shared by the group, that major should add one to the tally of major prizes won by insiders. The insider contributed one-tenth of the expenditure and should be credited with one-tenth of a win of a major prize.

If the data from the Fifth Estate surveys has not been discarded, it could be used to examine these other potential sources of uncertainty in more depth. Such an examination is likely to enhance the agreement between chance expectation and the actual wins by insiders, but it would require more time to conduct and lengthen this review report. The sources that have been addressed in depth are sufficient to establish plausibility for the assertion that there is no anomaly here, i.e. that the difference between the number of major wins by insiders and the expected number of major wins implied by the amount spent on the lottery by insiders is within the limits of normal chance variation.

Summary and Conclusions

The two major sources of uncertainty considered in this review are: uncertainty regarding the total number of insiders; and uncertainty concerning the adjustment factor to account for underreporting bias. The first of these sources accounts for the difference between Estimate #1 and Estimate #3. Estimate #1 uses 36,050 as the total number of insiders and Estimate #3 uses 101,174. The larger figure was based on "representative" samples from each of twelve types or channels of retail outlet. Rosenthal describes this figure as "inflated", and it could indeed be an overestimate. It could also be an underestimate. The uncertainty associated with estimates obtained from representative samples is extremely hard to assess analytically. That is a drawback not present when the sample is obtained via randomization. The smaller figure of 36,050 was obtained from a random sample instead of a representative sample, but it appears to be subject to a serious form of uncertainty for which the technical term is bias. The Rosenthal report implies that the sampling frame for the Fifth Estate survey consisted only of convenience stores. If so, 36,050 is almost certainly a seriously biased underestimate of the total number of insiders.

Estimate #1 is in effect doubly biased low because of underreporting. An attempt to account for underreporting was made by means of an adjustment factor estimated from a small survey of the general population conducted by telephone. When we use the larger count of insiders as a correction of the undercount bias in Estimate #1 and the adjustment factor from the small survey to correct for the underreporting bias, we get Estimate #3a. When the actual number of insider wins is compared to the expected number derived from Estimate #3a, we find that it is no longer "absolutely inconceivable" that the excess could occur by chance alone. The expected number of insider wins based on Estimate #3a does not

adequately reflect the uncertainties inherent in this analysis because it fails to incorporate the uncertainty associated with the adjustment factor for underreporting. That factor was estimated from the small additional survey and is as a result itself subject to uncertainty. The data for assessing uncertainty in the adjustment factor was not in the Rosenthal report, but an educated guess allowed a calculation of the impact it might have on the expected number of insider wins. That calculation brought within the realm of plausibility the possibility that the expected number of insider wins might even exceed the actual number. In other words, it is a reasonable possibility that insiders may not have won as many major prizes as they should have on the basis of the amount they spent on the lottery.

The conclusion here can be simply stated. When the various sources of uncertainty impacting the calculation of the expected number of wins by insiders are taken into account, it is reasonable to infer that the difference between that expected number and the actual number of wins by insiders may well lie within the limits of normal chance variation.

Donald S. Burdick
January 23, 2007

METAMETRICS, INC.
Bringing Meaning to Measurement

1000 Park Forty Plaza Drive, Suite 120, Durham, North Carolina 27713 • Phone (919) 547-3400 • Fax (919) 547-3401
E-mail: Webinfo@Lexile.com • Web site: www.Lexile.com