Insider Wins in the Ontario Lottery

1. Introduction. I have been asked to provide an evaluation of various statistical issues that have arisen recently concerning the expected number of insider winners in games managed by the *Ontario Lottery and Gaming Corporation* (indicated by OLG, hereafter). In particular, I was asked to respond to the following questions:

(i) are assertions made on TV in fact valid,

(ii) based on the available information what would be a reasonable number of expected retail winners and

(iii) are the number of retailer winners in fact disproportionately high or within an expected statistical range based on the frequency of play/spend versus the general population.

The remainder of this report is laid out as follows. In the next section I give some general comments concerning the statistical problem of estimating the expected number of insider winners and look in detail at the estimation technique that has been employed in the so-called "Rosenthal Report" (indicated by RR, hereafter). This examination reveals how the estimator depends on various unknown quantities related to the size, betting frequency and chances for winning of the insider population relative to that of the general public. Section 3 demonstrates how choices for these unknown parameters can produce results that suggest the observed number of insider winners ranges from anywhere between ordinary to actually being too small. I conclude in Section 4 with some general comments concerning the scientific merit of the work that has been done to date and suggest alternative questions that have not been considered but may be of more relevance in the context of insider winning.

2. What is being estimated? In reading through several documents concerning the number of insider winners for OLG I have frequently come across the term "Expected Number of Insider Winners." However, I have as yet to see a precise definition for this quantity. Instead, it seems to be defined as whatever the current number in the discussion might be estimating. This is not the way Statistics and, more generally, Science is done. The quantity being estimated *must* be defined first and only then does it becomes meaningful to consider strategies for its estimation.

There is, in fact, a well defined meaning for the word "Expected" in probability and statistics. It is the average value for the random variable in question (in this case the number of insider winners). In the current setting this translates into

$$\begin{aligned} &\text{Expected number of insider winners} \\ &\quad = \ (\text{number of insider plays}) \times (\text{probability of insider winning}). \qquad (1) \end{aligned}$$

So, assessment of the expected number of insider winners is conceptually relatively simple. It requires only the selection of two numbers: namely, i) the number of insider plays and ii) the probability of insider winning. Perhaps of more importance is that *any* estimation method that is being used for the expected number of insider winners (e.g., the one used in the RR) is implicitly making choices for these two numbers in some fashion.

Formula (1) provides the expected number of insider winners corresponding to any given number of insider plays for lottery games. One problem is direct assessment of the actual number of insider plays. In practice, this quantity has been viewed as being equivalent to dollars wagered. It is worth noting that since some instant games involve wagers of more than one dollar, the two quantities are not actually equivalent. This point seems to have been largely ignored.

Because estimation of the expected number of insider winners necessarily entails estimation of the the number of insider plays and the probability of insider winning, this raises questions about how such quantities are being treated in the estimators that have appeared in the RR and other sources. To simplify matters it will be helpful to adopt the following notation:

$$\begin{aligned} E_T &= \text{total number of expected winners,} \\ \hat{E}_T &= \text{observed number of winners,} \\ E_I &= \text{expected number of insider winners,} \\ N_I &= \text{number of insiders,} \\ P_I &= \text{probability of winning for an insider,} \\ S_I &= \text{average spending on lottery by insiders,} \\ N_P &= \text{population size,} \\ P_P &= \text{probablity of winning for all players and} \\ S_P &= \text{average spending on lottery by general population.} \end{aligned}$$

Then, RR type estimators emerge from the following string of formulae (the symbol $\approx$ signifies approximate equality in some heuristic sense):

$$
\begin{aligned}
E_I &= (\text{number of insider plays}) \times P_I \\
&\approx N_I \times S_I \times P_I \\
&= E_T \times \frac{N_I \times S_I \times P_I}{E_T} \\
&\approx E_T \times \frac{N_I \times S_I \times P_I}{N_P \times S_P \times P_P} \\
&\approx \hat{E}_T \times \frac{N_I}{N_P} \times \frac{S_I}{S_P} \times \frac{P_I}{P_P} \qquad (2)
\end{aligned}
$$

The value for $\hat{E}_T$ that has been used in formula (2) in the RR is 5,713 corresponding to the number of winners among the general public between 1999-2006. The quantity $\frac{P_I}{P_P}$ has been assumed to be 1 which corresponds to the (null) *hypothesis* that insiders have the same chance of winning or, more to the point, play essentially the same games with the same relative frequency, as other players. The value of N_P represents the population of Ontario and is known to be around 8.9 million. This leaves the values for N_I and the ratio $\frac{S_I}{S_P}$ which are replaced by various crude upper bounds based on some survey information and speculation. It we look beyond any difficulties that might arise from the *ad hoc* selection of N_I and $\frac{S_I}{S_P}$ there are also some fundamental technical problems associated with this approach to estimation that include the following:

(i) the quantity $\hat{E}_T$ is a random variable yet it is treated as a constant from an estimation error perspective and

(ii) there is no obvious way to assess the standard error for the estimator which limits its utility. Some *ad hoc* attempts to allow for estimation error have been used that replace the estimators of N_I and $\frac{S_I}{S_P}$ by one standard deviation upper bounds. This is not a standard approach, ignores item 1 and, in addition, overall bounds created in this way can have much higher error probability than that associated with each individual bound.

In my opinion, the use of formula (2) for estimation of the expected number of insider winners cannot be justified without, at the very least, the development of an associated standard error estimator. There are enough similarities with estimators of totals and ratios from sample survey work and capture-recapture studies that it may be possible to devise an effective estimator through study of that literature. Assuming such to be the case, estimation could then be accomplished using a sample survey whose results would allow for estimation of N_I and $\frac{S_I}{S_P}$.

An alternative and likely more fruitful endeavor would be to discard the current estimation strategies and, instead, consider other forms of statistical inference about the winnings for insiders that might have more practical relevance for OLG and it's players. The hypothesis being tested by the statistic produced from formula (2) is that $P_I = P_P$. It is unclear to me that this particular hypothesis is of any practical relevance. I will discuss this point further in Section 4.

3. Number games.

Formula (2) was used in RR to produce a variety of numbers that suggested the observed number of insider wins from 1999–2006 was in some sense "unlikely." In this section I will show how similar calculations based on supporting evidence can produce the opposite conclusion. The point of this exercise is not to promote one set of calculations over the other. But, rather to demonstrate that the outcomes can vary widely depending on the assumptions that are made.

My reading of RR suggests that the phrase "unlikely" in that particular context means that something as large or larger than the observed number of insider winners between 1999–2006 (i.e., 197) has a low probability of occurring if $\frac{P_I}{P_P} = 1$. This places the problem into a the realm of statistical hypothesis testing with formula (2) being used (with $\frac{P_I}{P_P} = 1$) to "estimate" what the expected number of winners should be under the null hypothesis that $\frac{P_I}{P_P} = 1$. An estimator of the chance of seeing 197 or more insider winners under the null model is then obtained by using this estimated expected value in conjunction with a binomial distribution with 5,713 trials (or an associated Poisson approximation to the binomial). This is, at best, an approximation (with unknown accurracy) since there is no obvious reason to presume that this particular binomial distribution will accurately represent the actual probability distribution for the number of insider winners. If we simply ignore such considerations, then the only problem that remains is the choice of N_I and $\frac{S_I}{S_P}$. Results from a survey of 380 retailers conducted by Research Dimensions estimates $\frac{S_I}{S_P}$ to be around 1.9 and they have evidence which suggests $N_I = 140,000$. The corresponding approximate upper tail probability or "P-value" that results from these figures is about 4% if one includes a crude upper bound (similar to that in RR) that allows for estimation of E_T by $\hat{E}_T$. If $N_I = 140,000$ is replaced by $N_I = 175,000$, for example, the "P-value" becomes 92.8% suggesting that there might even have been too few winners if there were that many retail employees.

Of course the key factor in the previous calculation is the assumption that $\frac{P_I}{P_P} = 1$. Since insiders are in direct and frequent contact with players that both win and lose it would be surprising if they learned nothing from such encounters in that their playing habits exactly parallel those of the general public. Once we relax this restriction then essentially anything becomes feasible depending on the value assigned to $\frac{P_I}{P_P}$. For example, using $N_I = 60,000$ as in RR with $\frac{P_I}{P_P} = 2.5$ gives a "P-value" of 22%. Similarly, with $N_I = 140,000$ and $\frac{P_I}{P_P} = 1.05$ the "P-value" is about 15%. In a very rough, entirely non-rigorous sense, the two results for the $N_I = 140,000$ case suggest that the data finds more support for the hypothesis that insiders win 5% more of the time than the general public than that they win equally often.

In summary, the answer to question (ii) of the introduction is that, in terms of the available information, one can provide arguments that support a very wide range of "expected number of insider winners" using the estimator employed in the RR. In addition, the fact that this estimator (or test statistic) has no known distributional properties means that it cannot, at present, be used in any meaningful way to provide an expected range for number of insider winners under its corresponding null hypothesis.

4. Conclusions.

The first question on the list in Section 1 concerns whether or not the claims in the media could be true. These claims are predicated to a large extent on information from a sample survey conducted by the Fifth Estate that was analyzed by Jeffrey Rosenthal. Professor Rosenthal is a distinguished scholar. His academic publication record in peer reviewed professional journals is impeccable. However, his report that appears on the Fifth Estate web site should not be confused with or given similar credence as his scientific publications. Instead, this article should, in my opinion, be viewed as a non-scientific, opinion piece or editorial. Reasons for this assessment include:

(i) absence of peer review,

(ii) failure to follow standard scientific principles concerning study design and objectives,

(iii) conclusions drawn from *ad hoc* calculations without supporting statistical distribution theory and/or standard error assessments and

(iv) excessive hyperbole that signifies a possible loss of objectivity.

I consider the calculations that appear in Professor Rosenthal's "report" to be similar to what I did in Section 3 of this document. That is, some quick and dirty number crunching to provide some initial insight into a problem under investigation. The purpose of such exploratory work is usually to aid in the formulation of pertinent scientific questions that can then be addressed through appropriate data collection and subsequent analysis using well established statistical methodology. Unfortunately, the next step in this process has as yet to occur and all that has been presented are the exploratory scribbles. These results are simply too *ad hoc* and too poorly formulated to be deemed conclusive in any sense about any point they might be trying to make.

I would now like to return to my comment in Section 2 concerning possible reevaluation of the entire "expected number of winners" issue. The emphasis that has been placed on estimating this particular measure is largely a consequence of Rosenthal's use of this quantity in his work. But, upon closer inspection, his use of "expected number of insider winners" can be recognized as an indirect route to address questions about the probability of insider winning and that is the real object of interest. His goal is then to test (i.e., attempt to reject) the hypothesis that the probability of winning for insiders is the same as that of the general public. Had this point been carefully considered the Fifth Estate "study" might never have gone forward and almost surely something of a different nature would have been undertaken. In short, adherence to the principles of the Scientific Method would certainly have been a cost-effective approach in this case.

So, let me conclude by asking the question "Why would one want to test the hypothesis that insiders and the general public win with equal relative frequency?" What would it mean if this hypothesis were false? Would this signify something ominous or merely that people (i.e., the insiders) are capable of learning from the experience of others? Of course the answer to the latter question is that it could be due to either factor and this simply means that one is testing the wrong hypothesis if separating out the two options is the objective. More relevant questions concerning the winning probabilities for insiders might include:

(i) Does observation of win/lose experiences for lottery players provide information that can be used to improve chances for winning and, if so, by how much?

(ii) Is the probability of winning among insiders consistent with the answer to question 1)?

Only in the case of a negative answer to question 2) would there be cause for concern. Unfortunately, the Fifth Estate survey and Dr. Rosenthal's "analysis" provide no insight into either of these two questions.