Examination of the Statistical Analysis Carried out by the CBC during an Investigation of Insider Lottery Wins on Television's The Fifth Estate *

Fred M. Hoppe †

Department of Mathematics and Statistics,

McMaster University, 1280 Main St. W.

Hamilton, Ontario, L8S 4K1.

January 29, 2007

*This work comprises an independent review carried out under contract for the Ontario Lottery and Gaming Corporation.

†Professor of Mathematics and Statistics. E-mail: hoppe@mcmaster.ca.

1 Objectives

CBC Television's The Fifth Estate carried out an investigation, the results of which were televised on October 25, 2006, of insider (defined to be lottery retail outlet employees, managers, and owners) wins as a result of an apparent disproportionate number of retail winners of major prizes ($50,000 or higher) in comparison with the general public. The number of retailer wins during the seven year period 1999 - 2006 was taken as 200 in the analysis out of approximately 5,713 such prizes. The CBC concluded that the probability that this many prizes could be won by retailers was vanishingly small ("one in a trillion, trillion, trillion, trillion"). The statistical analysis relied on some limited survey data and estimates of the size of the retailer cohort.

In order to better understand the data and its implications, the Ontario Lottery and Gaming Corporation asked me to carry out a critical review of the CBC's analysis to outline strengths and weaknesses, as well as any short comings, and to determine:

- if the assertions made on TV are valid.
- a reasonable number of expected retailer wins based on the available data.
- whether the number of retailer wins are in fact disproportionately high or within an expected statistical range based on the frequency of play/spend vs the general population.
- what additional data would be needed to further substantiate the number of expected wins vs. the number of actual wins.

The OLG provided me with their internal retailer research, sales data and winner information, and the report [2] containing the probabilistic analysis, also available on the CBC website [3].

My main findings are presented in the next section followed by explanations and details in the succeeding two appendices.

2 Main Findings

As described in Appendix A, the CBC calculation relied on the Poisson model. However, the basis for this model was not explained so I rederived it myself in order to understand the implications of any assumptions, the effect of changes in the estimated parameters, and the consequences of deviations from the model. It is shown in Appendix A that a Poisson model assumes an idealized universe in which all ticket choices are equally likely. In such a model retailers and the general public play similar games and the Poisson arises as an approximation to the hypergeometric distribution. It turns out that the critical parameter is the fraction of all tickets sold that are attributable to retailers, that is the ratio of the number of retailer ticket purchases divided by the number of general public purchases, denoted by the symbol r below. In [2] sales data for all ticket purchases was used to estimate the denominator of r, while sample survey data on frequency of retailer purchases compared to the general public, together with an estimate of the number of retailers, was used to estimate the numerator.

One problem with using sales data is that they are not directly proportional to ticket sales because of varying ticket prices. For instance, a Lotto Super 7 ticket allows three plays. Another difficulty is that the frequency of retailer purchases was obtained from a small sample and is subject not only to sampling error but to non-response bias and response bias, such as underreporting. In addition, the assumed size of the retailer base in the CBC was much smaller than what OLG states, although alternative bases were considered at the end of the CBC's report.

In order to understand the sensitivity of the P-values I posited different possible scenarios changing the size of the retailer base and the average yearly spending. The amounts used for average yearly spending were chosen to

be consistent with the cohort that entered into the base by accounting for different spending by owners, managers, and other employees using both CBC and OLG data. The P-values were found to be very sensitive to assumed input values. For instance, assuming a retailer base of 100,000 and a blended average yearly retailer expenditure based upon the value \$370 used by the CBC for the general and \$420 for the convenience channel[1] I determined an expected number of retailer wins of 177.4 and a P-value of 0.0503 using the methodology in [2]. The full range of possibilities is reported in Appendix A.

There is concern in my mind about the validity of the Poisson model. The fundamental premise in this model is that the winning tickets are uniformly distributed among all the tickets that have ever been purchased. However, lottery sales data show that this is not true. It may be possible to relax this assumption to rederive another model, which should be a mixture of Poissons, to account for the fact that some products (games) are bought (played) with higher frequency than others.

But for validity of such a more general model, the retailers would still need to play each game with the same frequency as the general population. Data on winners shows this is not true because there are some games with a large number of sales but having few winners among the retailers. One explanation is that retailers stay away from those games. The effect is the introduction of a bias which reduces the overall public winning rate in comparison to the retailers if retailers stay away from games that have a low frequency of winners in the entire population, but which the general public plays.

I am very concerned about the consequences of such differential playing. There is a phenomenon in statistics known as Simpson's Paradox in which a spurious correlation may appear when discrete data in 2×2 tables are pooled.

[1] From Research Dimensions survey and an underreporting factor of 1.77 used by the CBC

Using the retailer play frequencies in the CBC report, I present in Appendix B a hypothetical lottery with three types of games (the number of games can be increased) in which the combined totals arise from the actual data so that overall the retailers are winning at a higher rate than the general public, yet for each individual lottery, the general public is winning at a higher rate. This shows that Simpson's paradox is achievable for the actual lottery data and points to the dangers of pooling data over dissimilar lotteries.

My overall conclusion is that from the available data, the assertion that the retailers are winning at a higher rate cannot be justified because such an inference is dependent on the number of retailers and their yearly lottery expenditures, which are not known with a high degree of certainty. Additionally, P-values are highly sensitive to the input parameters and the small P-values obtained by the CBC are a function of the parameters they selected. These were based in part of a small sample survey that is subject to sampling error, response bias, and non-response bias, as well as an assumed size for the retailer base, and a postulated Poisson model. CBC also used average sales data rather than actual ticket sales which they aggregated over all games. Since not all lotteries are played with equal frequencies, this limits the applicability of the Poisson model.

As a result of differential choices of games between retailers and the general public the data could be manifesting a Simpson's paradox and the possibility exists, that the apparently large number of retailer wins is a manifestation of differential choices of games and higher playing frequencies in some games in comparison to the general public.

The kinds of data that would have been useful to have would be precise numbers of retailers' expenditures or tickets purchased for each lottery product individually, or at least a large sampling study. Such data have not been collected and it would be difficult to justify their collection. Moreover, I believe

that the atmosphere has been sufficiently poisoned since the CBC telecast, that future survey data would not be reliable.

3 Appendix A – The Poisson Model

All of the probabilities (P-values) cited in [2], derive from a Poisson model, which depends on a single parameter λ and which is estimated in that report indirectly through sales data and limited survey data. There is some difference between the OLG and the CBC over how λ should be estimated, and this matter is addressed below. In order to understand the components that enter λ and how accurate the Poisson is, I felt it is necessary to flesh out the details of the Poisson model, which is stated as a given. It derives from certain assumptions whose validity needs to be examined. The following is my derivation of the Poisson model which I believe would be similar to what underlies the use of the Poisson in [2].

There is a population of size N, which we may view as representing all tickets purchased during the seven years 1999 - 2006 for lotteries in which there was at least one major prize (defined to be one whose value is \$50,000 or more) among all possible prizes. Within this population there are m units representing the winning tickets where m is taken to be 5713.

A statistical model arises when we consider a simple random sample of size n, selected from this population. The value of n is the number of tickets that can be attributed to lottery retailers (owners, managers, and employees). Let X denote the number of winning tickets among the n selected (so X represents the retailer wins). The value of X is random so X is a random variable with a probability distribution and models the number of retailer winning tickets. By assuming that all selections of n tickets are equally likely, retailers and the general public are implicitly considered to have the

same chance of picking a winning ticket. If the distribution of X can be determined then the probability of obtaining a specified number of winning tickets or more, purely by chance, can be computed. This is the P-value.

The P-value is one way of measuring the weight of observational evidence against an assumed (null) hypothesis (in this setting, the null hypothesis is that any retailer and any member of the general public have the same chance of winning a major prize), in favour of an alternative hypothesis (in this setting the alternative is that a retailer has a greater chance of winning a major prize). It is defined as the probability of obtaining data that are as, or more, unusual that what was observed, relative to the alternative hypothesis. In the setting at hand, if there are too many retailer wins in comparison to what would be expected then the evidence points in the direction of a higher probability for retailer wins. The P-value is thus the probability that as many or more retailer wins could occur purely by chance. It is random, data dependent, and if very small represents an unlikely situation. In the language of significance testing one then rejects the null hypothesis that the retailers and general public have the same chance of winning, in favour of the alternative that the retailers have a higher chance of winning. This purely probabilistic statement is then *interpreted* by attributing this higher chance of winning as the result of fraud. This is the manner in which the CBC has used statistics in its analysis.

We can obtain the distribution of X by visualizing this population as being comprised of m green tickets (major prize winners) and $N-m$ orange tickets (non-winners of major prizes). The distribution of X is obtained by an experiment in which n tickets are randomly selected (akin to the selection of seven balls of 47 in Super 7, but with an enormous number of balls). X counts the number of green tickets in the sample. The distribution of X is exactly hypergeometric but can be approximated with a binomial distribution

on n trials and success probability $p \equiv \frac{m}{N}$, as long as p is close to zero, which is the case because p is the proportion of winning tickets among all tickets purchased. Moreover, if n is large, which is also true, since n is the number of tickets attributable to retailers, then this binomial can be further approximated by a Poisson random variable with mean

$$\lambda = np = n\frac{m}{N} = m\frac{n}{N} = mr$$

where

$$r = \frac{n}{N}$$

is the fraction of tickets purchased by retailers. As a result, it is not necessary to know n and N individually, only the ratio r (note that m is known – this is the number of major prizes won). Thus the computation of the P-value hinges on knowing the value of r and the assumption of the Poisson model.

The fundamental premise in this model is that the winning tickets are uniformly distributed among all the tickets that have ever been purchased. Under this assumption [2] the only difficulty in computing the P-value is that r is not known but must be estimated. The method used in [2] is to estimate r from the dollar amount of sales, instead of the actual numbers of tickets purchased. The latter statistics would be preferable but are difficult to obtain.

The value of r is estimated in [2] as

$$r = \frac{\$13,338,500}{\$2,220,000,000} = \frac{1}{166.4}$$

and multiplying r by $m = 5713$ leads to $\lambda = 34.33$, which is rounded up to 35 in the report. The denominator used in computing r comes from an

[2] This assumption groups all lottery games together. What can go wrong if this assumption is not valid, and retailers purchase tickets for games with different frequencies than the general public is considered in Appendix B.

estimated \$2,220,000,000 (\$2.22 billion) in total lottery sales per annum. The numerator is determined by multiplying an estimated number of retailers 36,050 by an average yearly per capita lottery expenditure by retailers of \$370. This figure \$370 was obtained from a small sample survey conducted by the CBC. Its website [3] did not give details of the survey, such as the type(s) of retailer channels nor separate data on owners, managers, or employees, which could be useful in extrapolating to the entire retailer base, nor indicate how the survey questions were phrased. Additionally, people generally do not keep a contemporaneous record of their yearly spending and may offer a figure that is widely off the true value. Besides, there may be both response bias and non-respons bias. In fact, both the CBC survey and a survey carried out by Research Dimensions for the OLG show that surveys underreport actual consumer spending. As there is a clear need to use an accurate figure when making a precise probabilistic computation, I will consider how changes in the estimated number of retailers and their spending affects the P-value.

I have a concern regarding the use of ratios of sales data in place of ratios of tickets bought. The two would be identical if each ticket cost the same, which is not the case. Retailers and the general public might be purchasing differentially different products with different prices. A Super 7 ticket costs \$2 and gives three sets of seven numbers – essentially equivalent to three tickets – while a Lotto 6/49 ticket also costs \$2 but only has one set of six numbers. It would seem more appropriate to weight the data in proportion to the frequencies with which the products were purchased.

A second concern is that even if the value of r could be correctly reconciled, the Poisson model assumes that X arises from a simple random sample of tickets. For this to be the case, it would be necessary that all products are purchased with the same frequency, or at least that retailers and the general public purchase the same products with the same frequency. Then

the Poisson model could be adapted to a weighted sum of Poissons.

In fact, the lottery data show that retailers and the general public purchase different types of tickets. For instance, among the offline games, during the seven years 1999 - 2006, BINGO generated $960,461,124 in sales and had the largest sales among all instant games. Yet over this period, there were only five claims by retailers, indicating that either retailers were extremely unlucky in this game, or they stayed away from it. The situation is similar with CFLIFE which had sales of $676,511,240 during this period but only six retail winners. by comparison, CWORD, with sales of $868,201,752 generated 28 retailer wins. The chance of winning in CWORD is 1/1,000,000 while the chance of winning in CFLIFE is 1/2,500,000 which is a ratio of 2.5:1. The expected number of CWORD winnings should therefore be 2.5 times the expected number of CFLIFE winnings. Multiplying the six wins in CFLIFE by 2.5 gives $6 \times 2.5 = 15$ equivalent expected wins taking into account different probabilities of winning. Now this is an average and adding two Poisson standard deviations to account for random fluctuations gives $15 + 2 \times \sqrt{15} = 15 + 11.62 = 26.62$, which is still less than the 28 retailer wins. (I am using the same Poisson approximation and 95% one-sided confidence as in [2].)

Differential playing also occurs among the online games. For example, PROPLS generated the third largest sales of such games in 1999 - 2006 (although it was unavailable in 1999 - 2000). However, there was only one retail winner. Similarly, although Super 7 had sales of $2,556,073,198, compared to Encore 6/49 sales of $723,396,744, yet among retailers, there were 28 Encore 6/49 winners compared to only 10 Super 7 retail winners. Since the dollar value of the sales reflects the number of tickets purchased, and the more tickets purchased for a product, the more winners expected, it is evident that retailers play games at different relative frequencies than the general public.

Failure to account for differential ticket purchases between retailers and the general public is a shortcoming of the simple Poisson model. Perhaps it can be accounted for in a more detailed analysis involving a mixture of Poisson, but it is not clear whether the required data would even be available.

3.1 Sensitivity of P-values to Model Parameters

The P-value is given by [3]

$$\mathbb{P}[X \geq 200|\lambda]$$

where the expected number of retailer winners is taken as $\lambda = 35$ in [2] based on a retailer base of 36,050 and an average yearly expenditure of \$370. The number 36,050 was estimated using an average of 10,300 lottery locations per year [4] which was multiplied by 3.5 (rounded up by 2.5 standard deviations) retailers per location as estimated from the CBC survey. The CBC's total of 36,050 seems more in keeping with the size of the convenience channel, and is in the same range as the estimate of 38,761 for the number of retailers in the convenience channel as reported by the OLG.[5] With regard to the value \$370 of yearly spending per retailer, that figure was also obtained from the CBC from their survey of convenience stores (rounded up approximately 1.4 standard deviations) taking into account those who refused to respond,

[3] In the CBC report, the number of retailer winners was taken to be 200 and that is the figure used below.

[4] This is consistent with OLG data ending October 2006 showing 7918 convenience channel locations and 2870 other locations for a total of 10,788 retail locations.

[5] Given the uncertainty in the figures available, the CBC report considered an alternate base of 60,000 as well as bases of size 101,174 and 140,217, provided by the OLG. The figure of latter includes anyone selling tickets, while the former removes the turnover portion of the retailers. The OLG believes that the numbers of retailers, including all employees, not necessarily only those who sell tickets, could be as high as 170,000 - 180,000. However, I have not used any numbers above 140,000 in this analysis.

and extrapolating to the entire population. Again, these numbers differ from the survey results[6] shown in Table 1 which found an average of $27.00 (with standard error of the mean of $3.90) to be the average amount reported by convenience channel retailers spent on lottery tickets during the four-week period preceding the survey (in fact, a weighted average to account for different numbers of owners, managers, and employees in the convenience channel). This works out to $477.75 per year (including 2.5 standard deviations) scaling by the number of weeks in a year – this is the only meaningful calculation without precise knowledge of actual amounts spent per week, which may vary depending on the jackpots.

Table 1: Spending by Retailers in the Convenience Channel

Retailer Type	Average Number of Retailers per Outlet	Average Spent in Previous Four Weeks
Owner	1.5	$44.10
Manager	1.5	$27.80
Employee	1.8	$12.30

I felt it important to examine the sensitivity of the P-values to different input parameters, such as the yearly expenditures and size of retailer base. The first calculation I did was to confirm the P-value in the Poisson model using $370 average yearly lottery expenditures assumed by the CBC. I considered different retailer bases and took into account the same same underreporting factor of 1.77 as the CBC. The results are shown in Table 2. The important parameter that determines the P-values is not just the assumed

[6]Table from Research Dimensions survey. This study also compared spending habits of the general public and found that retailers play upwards of 2 or more times what the general public plays.

Table 2: P-values and Estimated Expected Numbers of Wins for Different of Total Retailer Base – \$370 Yearly Expenditure

Retailer Base	36,050	100,000	110,000	120,000	130,000	140,000
P-value	0.0000	0.0149	0.1501	0.5720	0.9087	0.9924
$\mathbb{E}[X]$	60.7	170.5	185.4	202.2	219.1	236.0

retailer base, but the estimated expected number of retailer wins, denoted by $\mathbb{E}[X]$. Notice that when the retailer base changes by 10%, for instance between 100,000 and 110,000, the P-value changes by a factor of 10 or 1000% because of its highly non-linear behaviour.

Next, I considered the effect of using the survey data obtained by Research Dimensions. I took the middle value between \$370 and \$477.75 = \$423 for average yearly expenditure. The results are in Table 3.

Table 3: P-values and Estimated Expected Numbers of Wins for Different of Total Retailer Base – \$423 Yearly Expenditure

Retailer Base	36,050	100,000	110,000	120,000	130,000	140,000
P-value	0.0000	0.3679	0.8028	0.9832	0.9996	1.0000
$\mathbb{E}[X]$	69.5	194.9	211.9	231.2	250.5	269.7

The differences between Table 2 and Table 3 are considerable and it may not be appropriate to extrapolate the Research Dimensions data to the entire retailer base because their survey was restricted to the convenience channel[7]. An important consideration is that whatever cohort is used to define "retailer", it is necessary to take an appropriate weighted average of yearly spending for the respective group being included. Notice from Table 1 that the employees spend less than owners

[7]estimated to be 38,761

Table 4: P-values and Estimated Expected Numbers of Wins for Different of Total Retailer Base – Blended Yearly Expenditure

Retailer Base	36,050	100,000	110,000	120,000	130,000	140,000
P-value	0.0000	0.2040	0.5949	0.9175	0.9934	0.9998
$\mathbb{E}[X]$	78.4	188.2	203.1	219.9	236.8	253.6

or managers. To consider a retailer base of 100,000 or higher it therefore seems reasonable (with the objective being understanding how P-values change under different assumptions) to use the average yearly spending of \$477.75 for the retailers in the convenience channel and the CBC's lesser figure of \$370 for the remaining retailers, most of whom are employees, not managers or owners. The results are shown in Table 4.

Table 5: P-values and Estimated Expected Numbers of Wins for Different of Total Retailer Base – \$249 Yearly Expenditure in non Convenience Channel

Retailer Base	36,050	100,000	110,000	120,000	130,000	140,000
P-value	0.0000	0.0001	0.0023	0.0270	0.1436	0.4066
$\mathbb{E}[X]$	77.2	152.3	162.3	173.7	185.0	196.3

I also took the extreme case of using \$249 average yearly expenditure for all retailers not in the convenience channel. This is completely inconsistent with data indicating that retailers spend upwards of twice what the average adult spends[8] Nonetheless, I felt this was a useful computation in order to determine the effect of the size of the assumed retailer base. These results are shown in Table 5.

Finally, I was curious to learn how the P-value changed when the retailer base

[8]The CBC report assumed a factor of 1.5, while OLG data indicates this to be closer to 1.9.

was fixed at 100,000 and using a blended average lottery spending but varying the amounts attributed to retailers in the convenience channel. This is shown in Table 6.

Table 6: P-values and Estimated Expected Numbers of Wins for Different of Convenience Yearly Expenditures –100,000 Base

Convenience Yearly $	$400	$410	$420	$430	$440	$450	$460	$470
P-value	0.0278	0.0377	0.0503	0.0659	0.0849	0.1075	0.1340	0.1645
$\mathbb{E}[X]$	173.8	175.6	177.4	179.1	180.9	182.6	184.1	186.2

All these calculations, under various conditions, show the sensitivity of the P-value to the assumed parameters, many of which arise from small sample surveys. They are also computed under the assumptions of the Poisson model in which retailers and the general public play alike.

4 Appendix B – Simpson's Paradox

As pointed out in the previous section, data show that retailers appear to play games in different proportions than the general public. This negates the use of a Poisson model which assumes that all tickets are equally likely. But there is another feature related to grouping discrete data that arise from disproportionate strata.

The following example, constructed from summary data used by the CBC and also from the OLG, illustrates how a flaw in interpretation and inference may arise when discrete data are grouped. To illustrate, consider Table 7, which shows the number of general public wins and the number of retailer wins, as taken from [2], while the number of general public tickets came from OLG data. For the number

of retailer tickets, I took into account both the factor 166 and the alternate factor 100 from [2] and used 132, which is somewhere in the middle. The figure of 5,529,658,614 is consistent with the seven-year average per annum sales of $2.22 billion cited in [2] which works out to an average of $\frac{\$2,220,000,000\times 7}{5,572,000,000} = \$2.79$ per ticket which seems reasonable considering that games with major prizes of $50,000 or more cost at least $2. Table 4 is then an accurate representation of the structure assumed in [2].

Observe that the retailers are winning at a rate

$$\frac{200}{42,341,386} = \frac{1}{211,707}$$

while the general public are winning at a rate

$$\frac{5513}{5,529,658,614} = \frac{1}{1,003,022}$$

Both the general public and the retailers are doing very well in comparison to the actual probabilities of winning a major prize, which is between 1/1 million and 1/2.5 million for most of the games. The win rate for retailers appears excessively high as per the CBC's analysis. They are winning at a rate $\frac{1,003,022}{211,707} = 4.74$ times the rate of the general public, while in [2], from which the number 42,341,386 was computed, the retailers are winning at a rate between $\frac{200}{35} = 5.71$ and (the alternative) $\frac{200}{57} = 3.51$ so the win rate lies in between, as expected. (Note that I am not endorsing the CBC's results in this analysis, rather I merely assuming their numbers and exploring the consequences.)

Consider now a hypothetical province in which there are three types of lotteries, A, B, and C, in which the seven year data are given by the same Table 7 (again recall that this table derives from, and applies to, actual OLG data with the CBC estimates for the number of retailer tickets.) It was pointed out earlier that grouping disproportionate data can lead to problems. I will now illustrate this.

Tables 8, 9, and 10 present the corresponding #Won and #Tickets for each of the three lotteries in this province. They reflect one possible way in which the combined data over all lotteries in Table 7 might arise and cannot be discounted.

Table 7: All Lotteries Grouped

	# Won	# Tickets
General Public	5,513	5,529,658,614
Retailers	200	42,341,386
Totals	5,713	5,572,000,000

Table 8: Lottery A

	# Won	# Tickets
General Public	2,800	5,055,687,876
Retailers	3	6,319,610
Totals	2,803	5,062,007,486

Overall, from Table 7, the retailers outperform (as measured by the ratio of proportions of winning tickets) the general public by the ratio

$$\frac{200/42,341,386}{5,513/5,529,658,614} = \frac{0.00000472351}{0.000000996987} = 4.74 : 1$$

However, in Lotto A (Table 8), the general public outperforms the retailers by the ratio

$$\frac{2,800/5,055,687,876}{3/6,319,610} = \frac{0.000000553832}{0.000000474713} = 1.17 : 1$$

In Lotto B (Table 9), the general public outperforms the retailers by the ratio

$$\frac{1,500/252,784,394}{100/17,694,908} = \frac{0.00000593391}{0.00000565134} = 1.05 : 1$$

Finally, in Lotto C (Table 10), the general public also outperforms the retailers by the ratio

$$\frac{1,213/221,186,345}{97/18,326,869} = \frac{0.00000548406}{0.00000529278} = 1.04 : 1$$

Thus, in each lottery, the general public are winning at a higher rate than the retailers while overall, when all lotteries are combined, the rates dramatically

Table 9: Lottery B

	# Won	# Tickets
General Public	1,500	252,784,394
Retailers	100	17,694,908
Totals	1,600	270,479,301

Table 10: Lottery C

	# Won	# Tickets
General Public	1,213	221,186,345
Retailers	97	18,326,869
Totals	1,310	239,513,213

reverse and the retailers are winning at a higher rate, nearly five times as much. It is instructive to relate these tables to P-values which is an alternate way to express the previous four ratios.

In Table 7, all lotteries grouped, the P-value is the probability

$$\mathbb{P}[X \geq 200 | \lambda = 43.74] = 0.0000.$$

In Table 8, only Lottery A, the P-value is

$$\mathbb{P}[X \geq 3 | \lambda = 3.50] = 0.46.$$

In Table 9, only Lottery B, the P-value is

$$\mathbb{P}[X \geq 100 | \lambda = 104.67] = 0.65.$$

In Table 10, only Lottery C, the P-value is

$$\mathbb{P}[X \geq 97 | \lambda = 100.24] = 0.60.$$

The combined table is highly significant, but the individual tables are not.

The phenomenon is known as Simpson's Paradox in the statistical literature. It is well-known and not really a paradox, although the first time someone is exposed to this the natural reaction is to not believe the numbers. But this notion is dispelled once the tables are added and shown to be correct. Generally, this phenomenon occurs when the entries in the table are disproportionate. It only takes one such disproportionate lottery to cause a reversal of rates.

As discussed previously, the data show that retailers tend to stay away from certain games that the general public plays. As a result, the situation is ripe for the numbers computed by the CBC to, in fact, arise in this fashion. Therefore, without knowing how many tickets were actually purchased by retailers and the general public, it is impossible to preclude the possibility of Simpson's Paradox, even assuming the numbers used by the CBC.

4.1 An Actual Occurrence of Simpson's Paradox

Real examples of Simpson's Paradox exist and the tables do not need to be as extreme as above. Here is one such case that I sometimes use in my statisics courses at McMaster University. The context is a common topic of everyday interest, sports. Figure 1 is a composite screen snap from the NBA website on Tuesday, February 5, 2002 showing percentages ("statistics") for Vince Carter (then of the Toronto Raptors, currently with the New Jersey Nets) and Paul Pierce (then and still with the Boston Celtics). Carter and Pierce arrived in the NBA the same year and were selected closely together in the college draft.

2000-01 STATISTICS

									REBOUNDS									
G	GS	MPG	FGM-A	FG%	3PM-A	3P%	FTM-A	FT%	OFF	DEF	TOT	APG	STL	BLK	TO	PF	PTS	PPG
44	44	38.9	443-961	.461	93-218	.427	241-307	.785	95	136	231	3.8	65	35	94	126	1,220	27.7

2000-01 STATISTICS

									REBOUNDS									
G	GS	MPG	FGM-A	FG%	3PM-A	3P%	FTM-A	FT%	OFF	DEF	TOT	APG	STL	BLK	TO	PF	PTS	PPG
51	51	37.8	416-916	.454	82-215	.381	331-431	.768	71	244	315	3.0	92	37	175	152	1,245	24.4

Figure 1: Vince Carter vs. Paul Pierce, Feb. 5, 2002

Table 11: Vince Carter vs. Paul Pierce

Overall

	Made	Attempts	%
Vince Carter	777	1486	52.3
Paul Pierce	829	1562	53.1

Field Goals (2 points)

	Made	Attempts	%
Vince Carter	443	961	46.1
Paul Pierce	416	916	45.4

3 Points

	Made	Attempts	%
Vince Carter	93	218	42.7
Paul Pierce	82	215	38.1

Free Throws

	Made	Attempts	%
Vince Carter	241	307	78.5
Paul Pierce	331	431	76.8

The shooting percentages of both players can be extracted from Figure 1 and appear in Table 11. Overall Pierce has a higher shooting percentage when the data for field goals (two points), three point shots, and free throws are combined. Yet, individually, in each of these three categories, Carter has a higher shooting percentage, a remarkable "reversal of fortunes." This example does show that Simpson's Paradox is a real phenomenon.

The explanation for this apparent parodox here lies also lies in disproportionate number of attempts by both players in the three different categories. In particular, Pierce had 431 free throw attempts compared to Carter. Although Carter had a better shooting percentage in this category, this category was the highest percentae for both players and weighted Pierce's totals more, sufficiently to reverse the overall shooting percentage in Pierce's favour.

References

[1] Ontario Lottery and Gaming Corporation.
http://www.olg.ca/

[2] J. S. Rosenthal. Analysis of Insider Ontario Lottery Wins.
http://www.cbc.ca/fifth/luckofthedraw/documents/statsreport.pdf

[3] The Fifth Estate. Canadian Broadcasting Corporation.
http://www.cbc.ca/fifth/luckofthedraw/index.html