J Gambl Stud
DOI 10.1007/s10899-009-9145-9

ORIGINAL PAPER

Benford’s Law and Number Selection in Fixed-Odds Numbers Game

Mabel C. Chou · Qingxia Kong · Chung-Piaw Teo · Zuozheng Wang · Huan Zheng

Abstract In fixed-odds numbers games, the prizes and the odds of winning are known at the time of placement of the wager. Both players and operators are subject to the vagaries of luck in such games. Most game operators limit their liability exposure by imposing a sales limit on the bets received for each bet type, at the risk of losing the rejected bets to the underground operators. This raises a question—how should the game operator set the appropriate sales limit? We argue that the choice of the sales limit is intimately related to the ways players select numbers to bet on in the games. There are ample empirical evidences suggesting that players do not choose all numbers with equal probability, but have a tendency to bet on (small) numbers that are closely related to events around them (e.g., birth dates, addresses, etc.). To the best of our knowledge, this is the first paper to quantify this phenomenon and examine its relation to the classical Benford’s law. We use this connection to develop a choice model, and propose a method to set the appropriate sales limit in these games.

Keywords Benford’s law · Numbers games · Customer segmentation · Small-number phenomenon

M. C. Chou · Q. Kong · C.-P. Teo (✉) · Z. Wang
Department of Decision Sciences, NUS Business School, National University of Singapore, Singapore, Singapore
e-mail: bizteocp@nus.edu.sg

M. C. Chou
e-mail: bizchoum@nus.edu.sg

Q. Kong
e-mail: qingxia@nus.edu.sg

Z. Wang
e-mail: zuozheng@nus.edu.sg

H. Zheng
Management Science Department, Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai, China
e-mail: zhenghuan@sjtu.edu.cn

Published online: 30 July 2009

Introduction

Gambling is probably one of the oldest inventions in human history. In the ancient past, it was often organized around a fight between tribesmen. This ancient game of skill has proliferated into the different sports betting games that are now commonly played in many countries. Gambling can also take the form of a game of chance where the winners are determined via an external event—a toss of bones, whoever draws the short straw, and so on. In fact, such games are now routinely played at a national or state level, where players bet on which prize-winning numbers will be drawn using mechanical devices (cf. Lafaille and Simonis 2005).

In many countries, number lotteries have become a popular source of revenue for governments. In 2005, the Hong Kong Jockey Club paid close to HK$12.4 billion to the SAR government in betting duties and profits tax. This is close to 8.6% of the total tax collected by the Inland Revenue Department of Hong Kong that year. In the same year, the Singapore government took in S$1.05 billion from the gaming operators in betting duties, against a total tax revenue of close to S$17 billion. These games are also popular in the West. A recent survey by the licensed operator of the UK National Lottery, Camelot, found that as many as 69% of the adult population in Britain played the lottery in 2005–2006.[1] On the other hand, while there is no national lottery in the US, similar games are now played in more than 30 states in the country.

There are many ways in which number lottery games can be organized. In a parimutuel game, the players bet on the outcome of the draw of (random) prize-winning numbers, with the winner drawing a fixed portion of the total amount of bets received. The payout for the winners in such games depends on the total amount of bets received and the total number of winners. On the other hand, in a fixed-odds game, the winner receives a fixed payout for each winning wager, and the total payout for the winner is proportional to the amount of the wagers he makes in the game. For a fixed-odds game, the prize is fixed for each ticket, and hence the return for each player does not depend on how other players bet. However, the game operator now bears the risk of paying out a large sum in prizes if a very popular number is chosen as the winning number. Most game operators handle the risk exposure issue in their fixed-odds numbers games by imposing a liability limit on the sales of each number-all future bets on those numbers with accumulated sales hitting the limit will be rejected. This raises an associated question-how should a game operator set the liability limit? Note that this issue is particularly important to legalized game operator as a large chunk of their sales will have to be returned to the government as tax revenues at the end of each year. This prevents the operator from building up a large reserve to absorb the exposure risk.

Teo and Leong (2002) used the Markowitz model to argue that it is reasonable to use a common sales limit for all numbers/bet-types in the game. They exploited the design of a popular four-digit numbers game played in Singapore to demonstrate the benefits of risk pooling in the liability limits management system. However, they focused mainly on internal risk control mechanism and did not study the impact of external demand distribution (i.e., how players select numbers) on the game. Interestingly, this turns out to have a huge impact on the effectiveness of the risk control mechanism.

[1] The BBC news article on this can be accessed at http://news.bbc.co.uk/1/hi/uk/6174648.stm.

Small Number Phenomenon

There are numerous studies in the gaming literature on lottery numbers selection among the players. One group of studies (e.g., Simon 1999; Henze 1997; Haigh 1997, and Ziemba et al. 1986) focuses on the lotto games (where players compete to pick, for instance, six winning numbers out of 45), and has revealed many interesting behavioral patterns showing how the players select their numbers. The most striking conclusion from these studies is that the players do not select their numbers randomly; that is, not all numbers are chosen with equal likelihood, and there is a tendency to select "auspicious" numbers (for instance, the number 7 is routinely chosen by players in the game in the UK; numbers below 31 are more popular than numbers above 31, etc.). Table 1 shows the proportion of bets received on each number from 1–45 (ranked from highest to lowest proportions), in a 1996 powerball game played in the UK (Tijms 2007).

Another group of studies (Chernoff 1999; Halpern and Devereaux 1989) focuses on the numbers game (where the players compete to pick the winning 3-digit or 4-digit number), which is also known as Pick-3 or Pick-4 in many states in the US. Halpern and Devereaux (1989) also observed that players in Pennsylvania favor small numbers in the 3-digit numbers game, where the winning number is drawn randomly from among the numbers 000–999. They observed that the bet volumes decrease rapidly from numbers in the 100s to 400s, then slowly to the 900s. A similar phenomenon was also observed by Chernoff (1999) in his study of the 4-digit game in Massachusetts.

The sales data received on a particular draw in Pennsylvania was clearly presented in Halpern and Devereaux (1989), which allows us to quantify this phenomenon in the numbers games. Figure 1 shows the empirical distribution of the sum-of-three-digits statistic of the numbers chosen by the players in the Pennsylvania game. We compare the empirical distribution against the base case where all the 3-digit numbers are selected with equal probability (i.e., the uniform-choice model). Interestingly, the empirical distribution indicates a leftward shift from the base-case distribution, indicating a general preference for smaller digits in the number selections.

This empirical evidence indeed suggests that players favor small numbers. We call this the small-number phenomenon in the numbers game.

Explanations

Studies investigating cognitions of lottery ticket purchasers showed that people failed to recognize that each number on a ticket is independent of the others. For example, Ladouceur et al. (1996) showed that adults were more likely to select the "most random" perceived combinations, although in reality each ticket was as likely to win as the others. In addition, Langer (1975) asserted that factors in a chance situation which are typically associated with skill situations (such as choice, competition, and passive or active involvement) cause an individual to believe they have control over a situation that is completely governed by chance. Ladouceur et al. (1996) found that individuals who selected their own lottery ticket requested a larger sum of money in order to relinquish or sell back their ticket than those individuals who were randomly given a ticket (machine generated numbers). They concluded that participants who were able to select their own lottery ticket perceived their ticket as having a greater chance of winning and, as a result, assigned a higher monetary value to the ticket than individuals in the no-choice condition. Erroneous beliefs commonly held by adult gamblers were also identified in Hardoon et al. (1997) and Ladouceur and Walker (1996). Herman et al. (1998) studied the question as to

Table 1 Popularity of the 45 numbers in a 6/45 powerball game

Rank	Number	Proportion
1	7	0.036
2	9	0.033
3	5	0.033
4	3	0.033
5	11	0.031
6	12	0.030
7	8	0.030
8	4	0.029
9	10	0.029
10	2	0.029
11	6	0.028
12	23	0.027
13	13	0.026
14	22	0.026
15	1	0.026
16	25	0.026
17	15	0.025
18	21	0.025
19	17	0.024
20	16	0.024
21	26	0.024
22	14	0.024
23	24	0.024
24	27	0.023
25	19	0.023
26	30	0.023
27	18	0.022
28	31	0.02
29	28	0.02
30	29	0.02
31	20	0.019
32	33	0.018
33	35	0.016
34	32	0.015
35	40	0.015
36	34	0.014
37	42	0.014
38	36	0.013
39	41	0.013
40	44	0.012
41	39	0.012
42	45	0.012
43	43	0.012

Table 1 continued

Rank	Number	Proportion
44	38	0.011
45	37	0.01

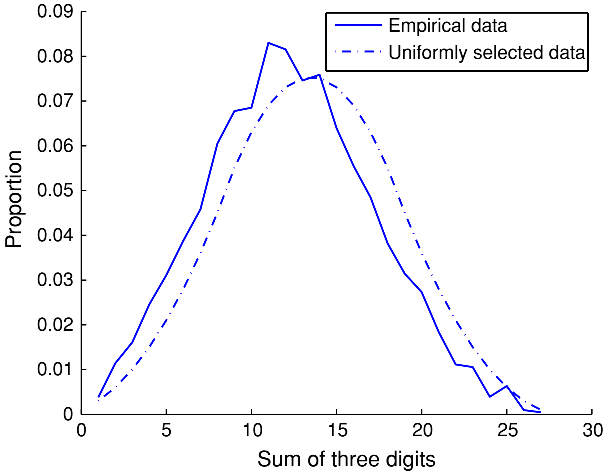


Fig. 1 Distribution of the sum-of-three-digits statistic in the 3-digit numbers game

when children's gambling behavior resembles that of adults. They showed that as children get older they are more specific in their beliefs that certain types of tickets are more likely to win than others.

There are a few explanations for the small-number phenomenon in lottery games. As stated in many studies (e.g., Halpern and Devereaux 1989; Simon 1999, etc.), a large proportion of players tend to select numbers associated with special dates (e.g., birthdays, anniversaries, etc.), meaningful numbers (e.g., phone numbers, car numbers, address numbers, etc.), and special events (e.g., accidents and murders), and these numbers tend to start with smaller digits (e.g., there are only 12 months in a year, so that the numbers 1–12 should be more popular than the numbers 13–45 in many 6/45 lotto games). Another explanation put forth by researchers is the observation that human beings simply can not choose numbers in a uniform manner. Loetscher and Brugger (2007) demonstrated using experimental methods that there is indeed a cognitive bias towards the selection of small numbers by human beings, even when they are told to select numbers "randomly." In one of their studies, a total of 488 subjects were told to "name a sequence of digits with each digit chosen from 1 to 6 as randomly as possible," and they found a surplus of small digits (1, 2, or 3) in all their experiments.

These studies, unfortunately, offer only anecdotal evidence (through surveys and interviews) and rudimentary explanations for the existence of the small-number phenomenon, and do not provide an analytical framework to quantify and model this phenomenon.

Another factor influencing the choice of numbers is superstitious beliefs, widely held by players of lottery games. In Chinese culture, certain numbers are believed by some to be lucky (or unlucky) based on the similarity of their pronunciation to that of certain Chinese words. For instance, Chinese people usually associate the digit 8 with prosperity, and thus numbers containing the digit 8 are normally more popular.[2] On the other hand, the number 4 is considered unlucky in many cultures in Asia, since it sounds like the word for "death" in spoken Chinese. Such beliefs concerning lucky and unlucky numbers tend to affect the popularity of certain numbers in the lottery game, leading to uneven distribution of the wagers on the different numbers.

Modeling Empirical Data

There have been several attempts to model the biases in the choice model of players. Simon (1999) considered the impact of the "lucky-number" biases and developed a model to approximate the distribution of the number of times that a combination would be chosen in the UK national lottery game. This provided a more accurate choice model for the UK lottery game, and fitted the data better than the uniform-choice model. Stern and Cover (1989) obtained the choice probability for each number in a lotto game, from the empirical marginal frequencies, by solving a related entropy-minimization problem. Ziemba et al. (1986) used regression methods and empirical data to estimate the popularity of each number combination in the lotto game. Haigh (1997) used choice probabilities directly on a set of numbers to estimate the popularity of the number combinations. Unfortunately, all these methods took the empirical data as given and focused merely on finding a better choice model to fit the empirical data. Thus, these methods did not exploit the existence of the small-number phenomenon in their modeling approaches, nor did they try to quantify this phenomenon.

Contributions

In this paper, we investigate the small-number phenomenon in the numbers games (rather than the lotto games) and use it to address the liability limits management problem. Our contributions in this paper are as follows:

- We quantify the small-number phenomenon through a curious fact observed by Newcomb (1881) and independently by Benford (1938). Interestingly, while the classical Benford's law captures the proportion of bets on the first significant digit reasonably well, it fails to account for the self-replicating nature of the empirical data beyond the first significant digit. By carefully modeling the ways players compose the digits in the numbers game, we refine Benford's law to develop an alternate consumer-choice model for different bet types using a handful of parameters only. Surprisingly, this parsimonious choice model is already able to capture some of the most important characteristics of the data in the numbers game.
- We examine the consequences of the small-number phenomenon on the prize liability performance of the game operator. In particular, our analysis suggests that it will be fruitful for the operators to pursue strategies to reduce the effect of the small-number phenomenon; that is, to promote or encourage players to choose numbers randomly. On the other hand, we show that the debate on whether a sales limit should be imposed on the game can be examined from the demand side—if numbers are selected in a uniform manner, then it may be futile to impose any sales limit since the performance is very sensitive to the total sales revenues of the game; that is, with a slight change in total

[2] In fact, China Mobile's Jiangxi branch held an auction to sell a "lucky" phone number recently, and one such number-with six consecutive eights—was sold for RMB 44,000!

sales revenues, the operator may swing from a situation with all numbers hitting the sales limit to a situation where all bets are accepted. Unless the total sales revenue can be accurately forecasted, it will be difficult to set the right sales limit in this environment. The small-number phenomenon in the choice process actually helps to stabilize this relationship between the total sales revenues and the proportion of numbers sold out. The imposition of a sales limit is thus more effective in such environment.

Modeling the Small-Number Phenomenon

Classical economic theory assumes that players behave like rational agents, and make decisions based on utility-maximization reasoning. As the returns from each number combination are identical, these players have no specific preference for any particular number and thus all numbers are selected with equal probability. We call these "Type 1" players.

However, recent empirical studies show that agents are not always seeking utility maximization in their decision making since framing, loss aversion, decision biases etc. can have major effects on players' decisions. To understand the small-number phenomenon, we need to augment the classical approach by incorporating the behavior of agents who pick their "lucky" numbers (arising from events in their daily life, or through superstitious beliefs) using reasoning which cannot be captured by any economic model. These players are superstitious, and have a general tendency to avoid betting on certain digits.[3] We call these the "Type 2" players.

We also assume that each player bets $1 on each number chosen.

Definition 1 Let β_B and β_N denote the proportions of type 2 and type 1 players respectively, with

$$\beta_B + \beta_N = 1. \tag{1}$$

The challenge in our problem is to estimate the proportions of type 1 and type 2 agents in the population of players based on the aggregate sales data. To this end, we need to have a better understanding on how type 2 agents choose their numbers. As these "lucky" numbers are normally selected from data series arising in the daily life of these type 2 agents, we will exploit a curious property associated with these natural numbers.

Benford's Law

Newcomb (1881) observed that the first few pages of books of logarithms were more worn than the last few and inferred that there might be more numbers starting with 1 or 2 than starting with larger numbers. Newcomb then drew a counter-intuitive conclusion that the first significant digits (i.e., first non-zero digits) of many data series in nature are not evenly distributed as expected, but follow a logarithmic law. Almost 50 years later, independently of Newcomb, Benford (1938) noticed the same phenomenon for categories of naturally occurring numerical data; for example, areas of rivers, atomic weights, numbers from Reader's Digest, and so on. He then came to the same conclusion, now known as Benford's

[3] Interestingly, our data suggests that players in Pennsylvania have an aversion to the digit 2, but favor digits 7 and 8.

law, which Newcomb had arrived at so many years previously. Both Newcomb (1881) and Benford (1938) proposed that the probability that a number has the first significant digit D_1 in a set [1…9], is given by

$$P(D_1 = d_1) = \log_{10}\left(1 + \frac{1}{d_1}\right), \quad \text{for all} \quad d_1 \in [1\ldots9]. \tag{2}$$

Let D_i denote the ith significant digit of a number. Hill (1995a) extended the above observation to a general version of Benford's law: for all $d_1 \in [1\ldots9]$, and $d_k \in [0\ldots9]$ for $k > 1$,

$$P(D_1\ldots D_i = d_1\ldots d_i) = \log_{10}\left(1 + \frac{1}{\sum_{j=1}^{i} d_j \times 10^{i-j}}\right). \tag{3}$$

Empirical evidence of Benford's law has appeared in a wide range of data; for example, stock index (Ley 1996), income tax (Nigrini 1996), mathematical series (Whitney 1972), and so on. Benford (1938) analyzed the underlying causes of this logarithmic phenomenon using a heuristic argument. Other mathematicians and statisticians have offered various explanations for this phenomenon. Raimi (1976) gave a review of some of the more intuitive explanations. It wasn't until 1995 that Hill (1995a) provided a formal rigorous proof that Benford's law is the only probability distribution which is scale-invariant and base-invariant. Using modern mathematical probability theory, and the scale- and base-invariant proofs, Hill rigorously demonstrated that the "distribution of distributions" given by random samples taken from a wide variety of different distributions in fact satisfies Benford's law (cf. Hill 1998).

One of the main applications of Benford's law is in fraud detection, under the hypothesis that fabricating data which conform to Benford's law is difficult. Recent empirical evidence shows that true accounting data sets conform very closely to Benford's law (Thomas 1989; Nigrini 1996). On the other hand, fabricated data rarely conform to Benford's law. Therefore, digital analysis based on Benford's law has been proposed as a new tool for fraud detection in recent years. Another application of Benford's law has been in the design of computers. Schatte (1988) devised rules that optimize computer data storage, by allocating disk space according to the proportions dictated by Benford's law, based on the assumption that input request satisfy Benford's law. Furthermore, both Varian (1972) and Hill (1995b) suggested using Benford's law as a test of the reasonableness of forecasts of a proposed model. If real life data follows Benford's law, it seems reasonable to assume that a good mathematical model should also do so.

In this paper, we add to this growing list of applications by showing that Benford's law can be used to capture the number selection behavior of type 2 agents in our model.

Choice Model for Type 2 Agents

WLOG, we will develop the choice model based on a 3D game, using the sales data published earlier in Halpern and Devereaux (1989). We have cross validated this model on other empirical data in several other number games, but unfortunately, due to the sensitivity of the data, we could not report the results here. For ease of exposition, we ignore the bets received for the number 000 from subsequent analysis; that is, we assume that none of the players will place a bet on the number 000. Using this assumption, the betting profiles of the type 1 players are drawn from a uniform distribution where all the numbers from 001 to 999 will have an equal chance of being selected. We focus next on the betting behavior of the type 2 players.

To ensure that the number selected has exactly 3 digits, we assume that the type 2 player may choose to compose a 3-digit number by padding the number he or she has chosen with leading zeros.[4]

Definition 2 Let γ_i denote the proportion of type 2 players who are betting on numbers with i significant digits.

By definition,

$$\sum_{i=1}^{3} \gamma_i = 1. \tag{4}$$

We first state a very simple consumer-choice model, where the classical Benford's law holds directly for the 3-digit numbers played.

Assumption 1 We assume that the type 2 player will choose to play the 3-digit number $d_1 \ldots d_i$ ($d_1 > 0$), with $3-i$ leading zeros, with probability

$$\gamma_i \log_{10}\left(1 + \frac{1}{d_1 \times 10^{i-1} + \cdots + d_i}\right). \tag{5}$$

Note that this is none other than the classical Benford's law, except that we weigh it with a factor γ_i to account for the proportion of players who bet with i significant digits.

It is now easy to prove the following proposition.

Proposition 1 Under Assumption 1, the expected proportion of the betting volume on a 3-digit number with first significant digit i, denoted by $E[S(i)]$, is

$$E[S(i)] = \beta_B \times \log_{10}\left(1 + \frac{1}{i}\right) + \beta_N \times \frac{1}{9}, \quad \text{for all} \quad i = 1, 2, \ldots 9. \tag{6}$$

Note that $E[S(i)]$ does not depend on γ_j. We can thus use this property to calibrate the value of β_B and β_N, by looking at the proportion of bets received for each significant digit. In the 3D data from Pennsylvania, the proportion of the type 2 and type 1 players are estimated to be 39.58% ($\beta_B = 0.3958$) and 60.42% ($\beta_N = 0.6042$), according to the least square model. We plot next the expected proportion of the first significant digit, given by the optimal parameter values, as shown in Fig. 2, along with the empirical proportion. The prediction from Benford's law captures the general trend in the empirical data, although we observe a general preference for first significant digit 3, 7 and 8 among the players, whereas the digit 2 has lower frequency than expected. Although we can refine our model to build in these biases into the model, we opted not to do so because such preferences do not appear to be universal across cultures.

While the simple model in Assumption 1 captures the behavior concerning the first significant digit rather accurately, we examine its ability to track the proportion of betting

[4] Note that this simplifying assumption may not hold in general, as some players may pad the numbers with trailing zeros, and some may simply duplicate the numbers to reach a 3-digit number. Halpern and Devereaux (1989) mentioned that triplets like 111 or 888 are very popular in the Pick-3 game in Pennsylvania. Unfortunately, it does not appear possible to incorporate such features into the model, without sacrificing the simplicity and tractability of the calibration model.

Fig. 2 Fitted proportion by the model

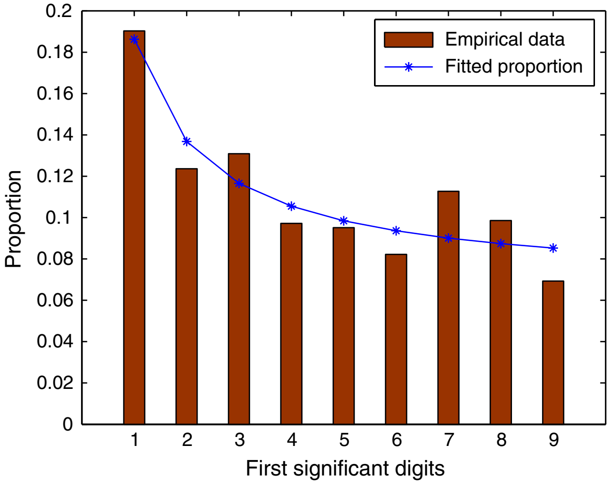
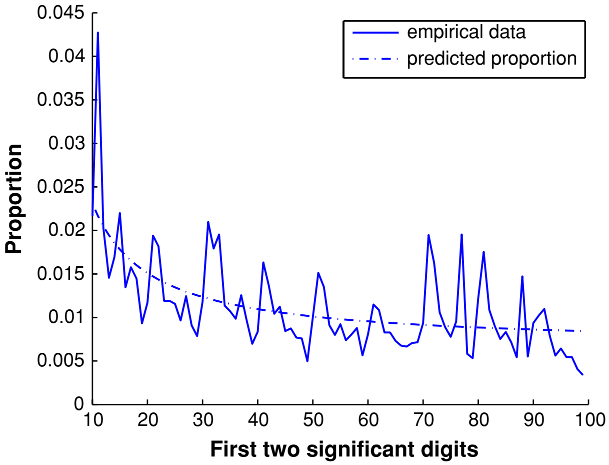


Fig. 3 Fitted proportion by the model for the first two significant digits



volume for the first two significant digits. We plot next the expected proportion of bets received and the empirical averages in Fig 3. Interestingly, our model is able to capture the declining popularity in the 3-digit numbers, as the first two significant digits grow from 10 to 99. This provides a partial explanation for the small-number phenomenon often observed in the games. Unfortunately, it could not explain the fact that the small-number phenomenon exists even in each decile (sub-block), as shown in Fig 3.

To understand the choice preferences beyond the first significant digit, we need to model an important characteristic in the way players compose the 3-digit numbers in the game. One such common strategy is to combine data from two different series to form a 3-digit number. For example, the number 246 could come from the 24th day of the month of June, or it could come from the address being level 6 of block unit number 24. The previous model assumes that the 3-digit numbers come from a single data series and hence fails to capture this switching behavior.

We notice that the probability distribution in our first assumption can be written in a different form:

$$\gamma_i \log_{10}\left(1+\frac{1}{d_1\times 10^{i-1}+\cdots+d_i}\right) = \gamma_i \log_{10}\left(1+\frac{1}{d_1}\right)\frac{\log_{10}\left(1+\frac{1}{d_1\times 10+d_2}\right)}{\log_{10}\left(1+\frac{1}{d_1}\right)}\cdots\frac{\log_{10}\left(1+\frac{1}{d_1\times 10^{i-1}+\cdots+d_i}\right)}{\log_{10}\left(1+\frac{1}{d_1\times 10^{i-2}+\cdots+d_{i-1}}\right)}.$$

Here, γ_i represents the probability that the type 2 player will pick a number with i significant digits. $\log_{10}(1+\frac{1}{d_1\times 10^{i-1}+\cdots+d_i})/\log_{10}(1+\frac{1}{d_1\times 10^{i-2}+\cdots+d_{i-1}})$ represents the probability that the *ith* digit is d_i, given that the first $i-1$ digits are $d_1 \ldots d_{i\text{-}1}$. To model the switching behavior, we refine the recursive approach in the following way:

- As before, $\log_{10}\left(1+\frac{1}{d_1}\right)$ represents the probability that the first digit is d_1.
- Let

$$\frac{\log_{10}\left(1+\frac{1}{d_1\times 10^{i-1}+\cdots+d_i}\right)}{\log_{10}\left(1+\frac{1}{d_1\times 10^{i-2}+\cdots+d_{i-1}}\right)+\lambda}$$

denote the probability that the player will continue to generate the *ith* digit d_i as if it comes from the same data series as the first $i-1$ digits, with parameter $\lambda > 0$. Note that in this way, the players will switch to a different data series with a non-negative probability

$$\frac{\lambda}{\log_{10}(1+\frac{1}{d_1\times 10^{i-2}+\cdots+d_{i-1}})+\lambda}.$$

- If the players switch to a different data series, let p_0 denote the probability that they will switch to the digit "0." Otherwise, they will switch to digit i, with $i \in \{1,\ldots,9\}$, with probability $(1-p_0)\log_{10}(1+\frac{1}{i})$.

With a slight abuse of notation, we can write

$$\log_{10}\left(1+\frac{1}{0}\right) := \frac{p_0}{1-p_0}, \quad \text{and} \quad \lambda := \frac{q}{1-q}.$$

We can now model the switching behavior in the 3-digit game in the following way:

Assumption 2 We assume that the type 2 player will choose to play the 3-digit number $d_1 \ldots d_i$ ($d_1 > 0$), with $3-i$ leading zeros, with probability

$$\gamma_i \log_{10}\left(1+\frac{1}{d_1}\right)\frac{(1-q)\log_{10}\left(1+\frac{1}{d_1\times 10+d_2}\right)+q(1-p_0)\log_{10}\left(1+\frac{1}{d_2}\right)}{(1-q)\log_{10}\left(1+\frac{1}{d_1}\right)+q}\times\cdots$$
$$\times\frac{(1-q)\log_{10}\left(1+\frac{1}{d_1\times 10^{i-1}+\cdots+d_i}\right)+q(1-p_0)\log_{10}\left(1+\frac{1}{d_i}\right)}{(1-q)\log_{10}\left(1+\frac{1}{d_1\times 10^{i-2}+\cdots+d_{i-1}}\right)+q}.$$

In this way, we can interpret the parameters as follows.

Definition 3 Let q denote the switching probability. Let p_0 denote the probability that the digit will be switched to 0.

Let $E[S(i,j)]$ denote the expected proportion of bets with first two significant digits i and j respectively.

Proposition 2 *Under Assumption (2),*

$$E[S(i)] = \beta_B \times \log_{10}\left(1+\frac{1}{i}\right) + \beta_N \times \frac{1}{9}, \text{ for all } i = 1,2,\ldots 9;$$

$$E[(S(i,j)] = \beta_B \log_{10}\left(1+\frac{1}{i}\right)\left(\frac{(1-q)\log_{10}(1+\frac{1}{i\times 10+j}) + q(1-p_0)\log_{10}(1+\frac{1}{j})}{(1-q)\log_{10}\left(1+\frac{1}{i}\right)+q}\right) + \beta_N\left(\frac{1}{90}\right).$$

Note that the expected proportion of first significant digits remains unchanged under both assumptions. The parameters under Assumption 2 are calibrated to be q = 0.9105, p_0 = 0.1054, to best fit the empirical data under the least square model.

We compare the expected frequencies of first two significant digits with those in empirical data respectively in Fig. 4. The frequencies generated from this model closely fit the frequencies of the empirical data. More interestingly, this model is able to capture the small-number phenomenon in the second significant digit of the data series.

Note that so far the parameters γ_j did not feature in the analysis. This arises because we have fixed the number of significant digits. To complete our specification of the choice model, we need to estimate the values of these parameters. Let $\hat{\gamma}_j$ denote the sample average of the proportion of bets with exactly $3-j$ leading zeros. We use $\hat{\gamma}_j$ to obtain an unbiased estimator of γ_j, using the following relationship:

$$\hat{\beta}_B\gamma_j + \hat{\beta}_N\frac{9\times 10^{j-1}}{999} = \hat{\gamma}_j, \quad \text{for} \quad j \geq 1. \tag{7}$$

Model Validation

We show next that the choice model under Assumption 2 proposed in the earlier section has the ability to track some of the most important characteristics of the betting data in the 3D game.

We first estimate the behavior for the sum-of-digits statistic, using data simulated according to Assumption 2 (with the calibrated parameters). We compare it with the empirical data (after removing the betting volumes on the 3-digit number 000). Figure 5

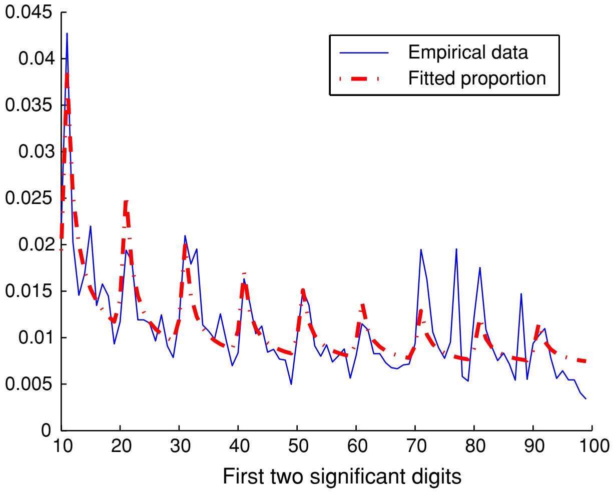
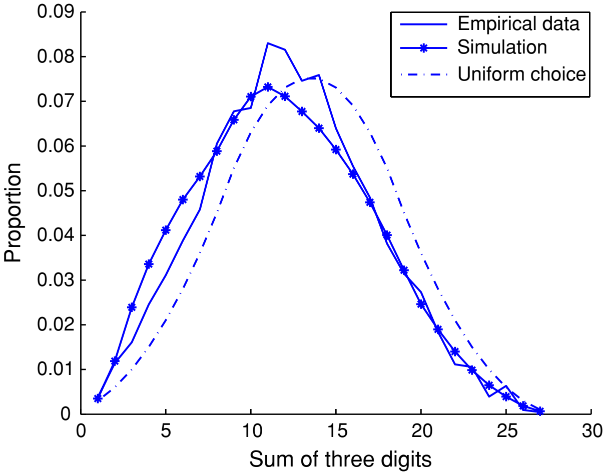


Fig. 4 Fitted proportion for the first two significant digits

Fig. 5 Distributions of sum-of-digits in empirical data, simulated data with Assumption 2, and uniform choice

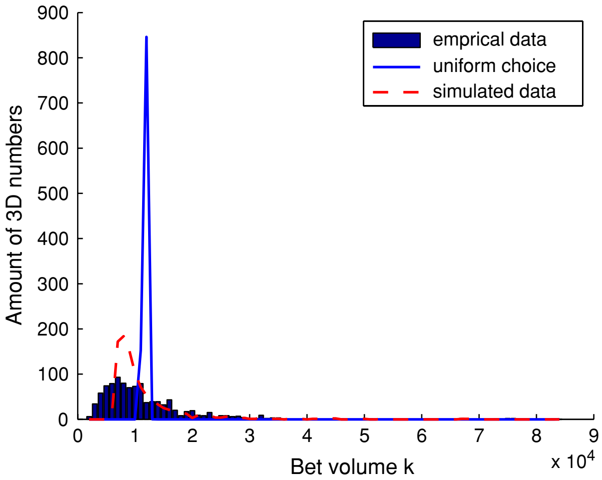


depicts the distributions of the sum-of-digits in three data series: the actual data, simulated data from our choice model, and the uniform-choice model.

The estimation of 39.58% type 2 and 60.42% type 1 players in the population seems right, as it captures the magnitude of the leftward shift in the empirical data reasonably well. Also, note that our choice model does not account for the superstitious beliefs observed in the empirical data (players generally avoid 2 and prefer 7 and 8). This partially explains why the proportions from our model are higher for smaller sum-of-digits (from 3 to 7) and lower for sum-of-digits around 15.

We track the performance of another statistic—the numbers of 3-digit bets attaining a certain betting volume. Figure 6 shows the number of 3-digit numbers in the game with a given betting volume (specified in the horizontal axis). The distribution obtained from the uniform-choice model follows a binomial distribution, and centers mainly around the mean. This yields a poor fit for the empirical data. The distribution obtained from our choice model clearly has a better fit.

Fig. 6 Betting volume of actual data, simulated data with Assumption 2, and uniform choice



Applications

The small-number phenomenon clearly has important implications for the operational risk management of the game. The numbers picked by the type 2 players introduce variability and skewness to the distribution of bets on the 3-digit numbers. The winning numbers, on the other hand, are randomly (i.e., uniformly) rolled out by a mechanical device, which implies that the hot numbers are chosen with the same probability as other numbers. The mismatch between the winning number distribution and the betting volume distribution leads to a significant operational risk: the operators may face a substantial payout if a popular number happens to be picked as the winning number! This is a phenomenon which often worried the game operators. In Quebec, according to Lafaille and Simonis (2005), "the first drawing caused a prize liability well in excess of the amount received in sales." Fortunately, "over the long run it all evened out and the projected prize percentage was achieved."

We show in this section that the small-number phenomenon plays a significant role in the large volatility of prize liability experienced by many operators in the game. We further exploit this observation to propose a method to help determine the sales limit in these games.

Volatility of Prize Liability

We examine the impact on the prize payout volatility by the proportion of type 2 players in the population of players. We compare the variability of payout in the 3D game, as we increase the proportion of type 2 players from 0 to 39.58% in the choice model (both with Assumption 1 and Assumption 2). For this study, we assume that the sales limit is higher than demands, so that all bets are accepted.

Consider a game with a prize P and N players, each betting \$1 on a number drawn from a respective distribution. Let $X_{\beta_B}(n)$ denote the amount of bets received on the number n when the proportion of type 2 players is equal to β_B. When the winning number for that prize is drawn uniformly among the 999 numbers (from 001 to 999, as we have ruled out the bets on the number 000), the expected payout in our choice model is simply

$$\frac{P}{999}\sum_{n=1}^{999} E(X_{\beta_B}(n)) = \frac{P}{999} \times N.$$

The second moment of the payout is

$$P^2\left(\frac{\sum_{n=1}^{999} E(X_{\beta_B}^2(n))}{999}\right).$$

Hence, the variance of payout is

$$P^2\left(\frac{\sum_{n=1}^{999} E(X_{\beta_B}^2(n))}{999} - \frac{N^2}{999^2}\right).$$

If all the N players choose their numbers independently, $X_{\beta_B}(n) \sim Bi(N,\ p_{\beta_B}(n))$, where $p_{\beta_B}(n)$ denote the probability that number n is picked in our choice model, given that the proportion of type 2 players is β_B. Hence,

$$E(X_{\beta_B}(n)^2) = N^2 p_{\beta_B}^2(n) + N p_{\beta_B}(n)(1 - p_{\beta_B}(n)).$$

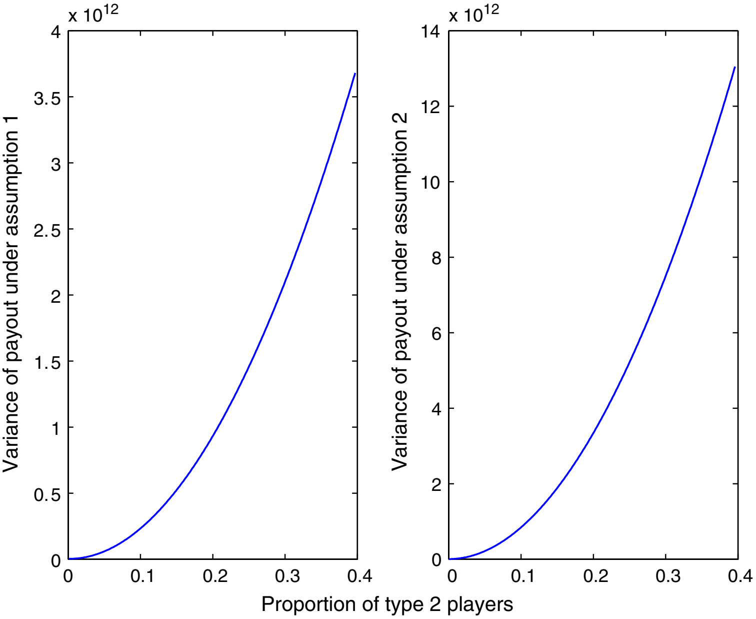


Fig. 7 Variance of payout as the proportion of type 2 players increases

We can thus analytically compare the variance of the payout, under different values of β_B. As shown in Fig. 7, under both assumptions, the variability of payout is increasing as the proportion of type 2 players increases. When β_B is equal to 0, that is, the demand is evenly distributed, the variance of payout is only 0.003×10^{12}. When β_B increases to 39.58%, the variance of payout under Assumption 1 is 3.6794×10^{12}, about 1216 times higher than that of the uniform-choice model. Since Assumption 2 captures more of the volatility of the data, the variance of payout is 13.049×10^{12} in this model, 4313 times bigger than that of the uniform-choice model.

Note that our choice model only includes 39.58% type 2 players and does not account for other random effects such as date or month effect. So, the volatility of actual prize payout should be even worse. For the 3D game in Pennsylvania, we conclude that the standard deviation of the prize payout can be reduced by 65 times if the proportion of the type 2 agents (β_B) reduces to 0.

Liability Limit

In the rest of this section, we use the small-number phenomenon to set the appropriate liability limit for the 3D game. Let D_n denote the (random) demand of a 3-digit number n. The distribution of D_n depends on the proportion of type 1 and type 2 players in the game. Let C denote the corresponding sales limit. Let S_n denote the accepted sales for number n; i.e.,

$$S_n = \min(D_n, C).$$

Note that

$$E[S_n] = C \cdot P(D_n > C) + E(D_n|D_n \leq C) \cdot P(D_n \leq C).$$

Let $R(S_1,\ldots, S_N)$ denote the "risk exposure" when sales for the N numbers are given by $(S_1,\ldots, S_N)$. There are several ways to model the risk measure $R(\cdot)$, and it generally depends on the distribution of the winning numbers drawn.

Suppose the expected return given \$1 bet is r. We use the mean-risk trade-off to model the utility function of the game operator. The expected utility function of the game operator is thus given by

$$r\sum_{n=1}^{N} E[S_n] - \lambda E\{R(S_1,\ldots,S_N)\},$$

where λ is an exogenous penalty term for risk exposure.

We can find C by solving the following maximizing problem:

$$\max_{C>0} \; r\sum_{n=1}^{N}[C\cdot P(D_n > C)+E(D_n|D_n\leq C)\cdot P(D_n\leq C)] - \lambda E\{R(\min(D_1,C),\ldots,\min(D_N,C))\}.$$

It can be easily shown that the objective function is convex. Thus, according to the first order condition, the optimal liability limit C satisfies:

$$\sum_{n=1}^{N} P(D_n > C) = \frac{\lambda}{r} E\left[\frac{\partial R(\min(D_1, C),\ldots,\min(D_N, C))}{\partial C}\right]. \tag{8}$$

Note that the left hand side corresponds to the expected number of hot numbers, i.e., the expected number of bet types reaching the sales limit in the draw. The sales limit can be set by merely choosing a sales limit C to control the number of hot numbers.

Suppose the total bets collected are to the value of \$$N$, and the cut-off limit is \$$C$ for each number. We next estimate the expected number of hot numbers (i.e., the numbers with betting volumes hitting the liability limit).

We define an indicator function $Y_{\beta_B}(n)$ as follows:

$$Y_{\beta_B}(n) = \begin{cases} 1 & \text{if} \quad X_{\beta_B}(n) \geq C; \\ 0 & \quad \text{otherwise.} \end{cases}$$

The expected number of hot numbers with liability limit \$$C$ is

$$\boldsymbol{E}\left(\sum_{n=1}^{999} Y_{\beta_B}(n)\right) = \sum_{n=1}^{999} \boldsymbol{P}\left(X_{\beta_B}(n) \geq C\right) = \sum_{n=1}^{999}\left(1 - \sum_{i=0}^{C-1}\binom{N}{i}(p_{\beta_B}(n))^i(1-p_{\beta_B}(n))^{N-i}\right)$$

Note we can use a normal distribution $N(Np_{\beta_B}(n), \sqrt{Np_{\beta_B}(n)(1-p_{\beta_B}(n))})$ to approximate the binomial distribution $Bi(N, p_{\beta_B}(n))$, if N is large enough. Hence, we have

$$\boldsymbol{E}\left(\sum_{n=1}^{999} Y_{\beta_B}(n)\right) = \sum_{n=1}^{999}\left(1 - \Phi\left(\frac{C - Np_{\beta_B}(n)}{\sqrt{Np_{\beta_B}(n)(1-p_{\beta_B}(n))}}\right)\right).$$

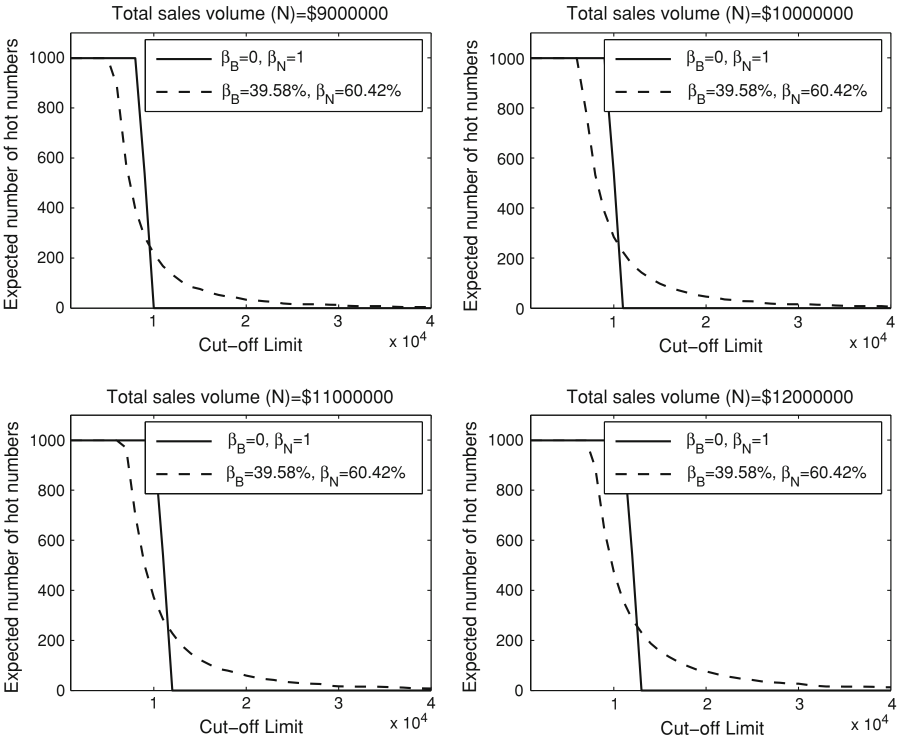


Fig. 8 Expected number of hot numbers using different liabilities

We can thus analytically compute the expected number of hot numbers given a liability limit $C, and compare the results using different liability limits. Figure 8 shows the expected number of hot numbers under different liability limits in the case that 39.58% players are type 2 and 60.42% players are type 1, and in the ideal case that all players are type 1 agents.

In the ideal case, because all numbers are selected with equal probability, the concentration of measure phenomenon kicks in and the expected number of hot numbers goes through a phase transition—dropping sharply from 999 (all sold out) to 0 (none sold out) for a narrow range of sales limit. This is most evident from Fig. 8: when the total sales is $9M, $10M, $11M, and $12M, respectively, the expected number of hot numbers drops sharply to zero when the liability limit is around $10000, $11000, $12000, and $13000, respectively. In this environment, trying to find the appropriate sales limit to control the right level of hot numbers is almost impossible because this number depends critically on the total sales level, a number which normally fluctuates from draw to draw.

In the empirical sales data, we have $\beta_B \approx 0.3958$. In this environment, interestingly, the phase transition phenomenon disappears, and the relationship between the sales limit and the expected number of hot numbers is more stable. For a sales limit of $10000, the hot numbers fluctuate from 200 to 400 when the total sales level changes from $9M to $12M. The relationship between the total sales and proportion of hot numbers are thus more stable.

Conclusion

In this paper, we have analyzed an interesting phenomenon in a popular numbers game. While it is by now folklore that players in these games prefer small numbers, this paper is arguably the first to quantify this behavior using Benford's law. The connection is forged by virtue of the argument that many natural data series satisfy Benford's law. We also take into account the choice behavior of the players, in particular the way the players compose the 3-digit number to obtain a refined choice model. Although we do not model additional phenomenon such as superstitious beliefs and date-month effect on the choice behavior, the simple model we built, using only a few parameters (i.e., the proportion of type 2 agents β_B, the probability of switching q, and the probability of padding the number with digit zero p_0), is already able to capture some of the most important characteristics of the empirical data.

While we have presented only the analysis using a set of publicly available data from the US, we have tested the model on an extensive series of data provided by a game operator in another region. Despite the differences in culture and beliefs, we found that the same underlying model can be used to describe the behaviour of the aggregate data, with the main difference coming from the proportion of Benford-like players. We believe that the small-number phenomenon is a generic behavior inherent in many numbers games.

The proportion of type 2 players (β_N) has a tremendous impact on the variability of the prize liability, and to a certain extent affects the appropriate choice of sales limit in the numbers game. There are many ways to mitigate the small-number effect through demand shaping. One approach, already in use, is to use on-site computer terminals to help players to pick the numbers randomly. However, people often do not like random picks because they like to assume some control over the outcomes (cf. Langer 1975). Therefore, there may be limits on the extent to which the industry can encourage random picks. Another approach is to re-design the game to encourage the players to bet on as many different permutations as possible. In Singapore, the introduction of a new iBet system (cf. http://www.singaporepools.com.sg) proves to be popular with the players. The new system allows the players to spread a dollar bet on as many permutations of the number combination as possible, with a corresponding reduction in the prize monies. It also helps to mitigate the effect of the small number phenomenon. Other possible approaches include posting the results of past winning numbers in the retail outlets to influence the selection of numbers by the players. The past winning numbers are drawn in a random manner and thus will not exhibit the same feature as numbers picked by Benford's type player. Of course, recency bias may actually deters players from betting on recent winning numbers, and hence this approach may not be as effective in persuading players from moving away from their preferred numbers.

Interestingly, if the game operators are able to reduce the proportion of benford players, then the above analysis shows that the imposition of the sales limit may no longer be needed, since it will be difficult and futile to implement such a mechanism anyway.

References

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society, 78*, 551–572.

Chernoff, H. (1999). How to beat the Massachusetts number game: An application of some basic ideas in probability and statistics. *The Mathematical Intelligencer, 3*, 166–175.

Haigh, J. (1997). The statistics of national lottery. *Journal of the Royal Statistical Society. Series A , 160*(2), 187–206.

Halpern, A. R., & Devereaux, S. D. (1989). Lucky numbers: Choice strategies in the Pennsylvania daily number game. *Bulletin of the Psychonomic Society, 27*(2), 167–170.

Hardoon, K., Baboushkin, H., Gupta, R., Powell, G. J., & Derevensky, J. L. (1997). *Underlying cognitions in the selection of lottery tickets*. Paper presented at the annual meeting of the Canadian Psychological Association, Toronto, Ontario, June.

Henze, N. (1997). A statistical and probabilistic analysis of popular lottery tickets. *Statistica Neerlandica, 51*(2), 155–163.

Herman, J., Gupta, R., & Derevensky, J. L. (1998). Children's cognitive perceptions of 6/ 49 lottery tickets. *Journal of Gambling Studies, 14*, 227–244.

Hill, T. (1995a). Base-invariance implies benford's law. *Proceedings of the American Mathematical Society, 123*, 887–895.

Hill, T. P. (1995b). A statistical derivation of the significant-digit law. *Statistical Science, 10*, 354–363.

Hill, T. P. (1998). The first digit phenomenon. *The American Scientist, 10*(4), 354–363.

Ladouceur, R., Dube, D., Giroux, I., Legendre, N., & Gaudet, C. (1996). Cognitive biases and playing behavior on American roulette and the 6/49 lottery. Unpublished manuscript. Québec: Universite Laval.

Ladouceur, R., & Walker, M. (1996). A cognitive perspective on gambling. In P. M. Salkovskis (Ed.), *Trends in cognitive and behavioral therapies* (pp. 89–120). New York: Wiley.

Lafaille, J. M., & Simonis, G. (2005). Dissected re-assembled: An analysis of gaming.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology, 32*, 311–328.

Ley, E. (1996). On the peculiar distribution of the U.S. stock indexes' digits. *American Statistician, 50*, 311–313.

Loetscher, T., & Brugger, P. (2007). Exploring number space by random digit generation. *Experimental Brain Research, 180*, 655–665.

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematician, 4*, 39–40.

Nigrini, M. J. (1996). A taxpayer compliance application of benford's law. *Journal of the American Taxation Association, 18*, 72–91.

Raimi, R. (1976). The first digit problem. *The American Mathematical Monthly, 83*, 887–895.

Schatte, P. (1988). On mantissa distributions in computing and benford's law. *Journal of Information Processing and Cybernetics, 24*, 443–445.

Simon, J. (1999). An analysis of the distribution of combinations chosen by UK national lottery players. *Journal of Risk and Uncertainty, 17*(3), 243–276.

Stern, H., & Cover, T. M. (1989). Maximum entropy and the lottery. *Journal of the American Statistical Association, 84*, 980–985.

Teo, C. P., & Leong, S. M. (2002). Managing risk in a four digit number game. *SIAM Review 44*(4), 601–615.

Thomas, J. K. (1989). Unusual patterns in reported earnings. *The Accounting Review, 64*, 773–787.

Tijms, H. (2007). *Understanding probability: Chance rules in everyday lives*. Cambridge: Cambridge University Press.

Varian, H. R. (1972). Benford's law. *The American Statistician, 26*, 65–66.

Whitney, R. E. (1972). Initial digits for the sequence of primes. *American Mathematical Monthly, 79*, 150–152.

Ziemba, W. T., Brumelle, S. L., Gautier, A., & Schwartz, S. L. (1986). *Dr. Z's 6/49 Lotto Guidebook*. Vancouvrr and Los Angeles: Dr. Z. Investments. Inc.