Powerball: A Wealth of Data for a Variety of Student Projects

Edward R. Mansfield
University of Alabama, Department of Information Systems, Statistics, and Management Science, Box 870226, Tuscaloosa, AL 35487-0226

Abstract

Be a winner with Powerball! Use their data to expand the experience and creativity of your students in analyzing data. Since the lottery game Powerball started in 1992, changes have been made that have altered the probability of winning the grand prize. The basic process of drawing five white balls from one bin and one red ball from a different bin has remained constant, but changes in the number of balls in each bin has occurred. Data for all drawings is readily available on the internet and also provided by the author in an edited form. Some activities: Ask students to determine which numbers are more likely or less likely to occur. Obviously all numbers are equally likely, but students may find some "hot" and "not" numbers due to changes in the game rules over time and by their looking only at default histograms. Examine the distribution of discrete order statistics. Model the changing size of the jackpots over time. Determine the breakeven point for the jackpot. Analogies are given to help one understand how *unlikely* it is to win the big prize.

When additional numbers are added to alter the probability of winning, the higher numbers will have lower frequencies when compared to older data because they have been available for fewer drawing. The reason may not be obvious to some students. Can the distribution of the discrete order statistics be determined from the data? How should the parameters of the distributions be estimated?

After a jackpot is won, the value is the payout resets to a base amount and grows each week based on the number of tickets sold. Is the rate of growth a function of factors such as the previous jackpot value, or the time of year, or other factors?
.

Key Words: Student projects, Data analysis**,** Randomness, exponential growth, applied discrete order statistics

1. Learning Objectives

The data sets available through this paper are useful for helping students improve their skills at researching on the internet the historical perspective for specific topic, examining patterns in data, describing data distributions by using graphs, observing the distributions of discrete order statistics, determining equations for exponential growth then seeing if growth patterns are the same, and assessing the merits of an investment opportunity.

Four data sets are available at **www.cba.ua.edu/~emansfie.** They contain the Powerball numbers arranged in different formats for different types of analysis.

2. Powerball Background

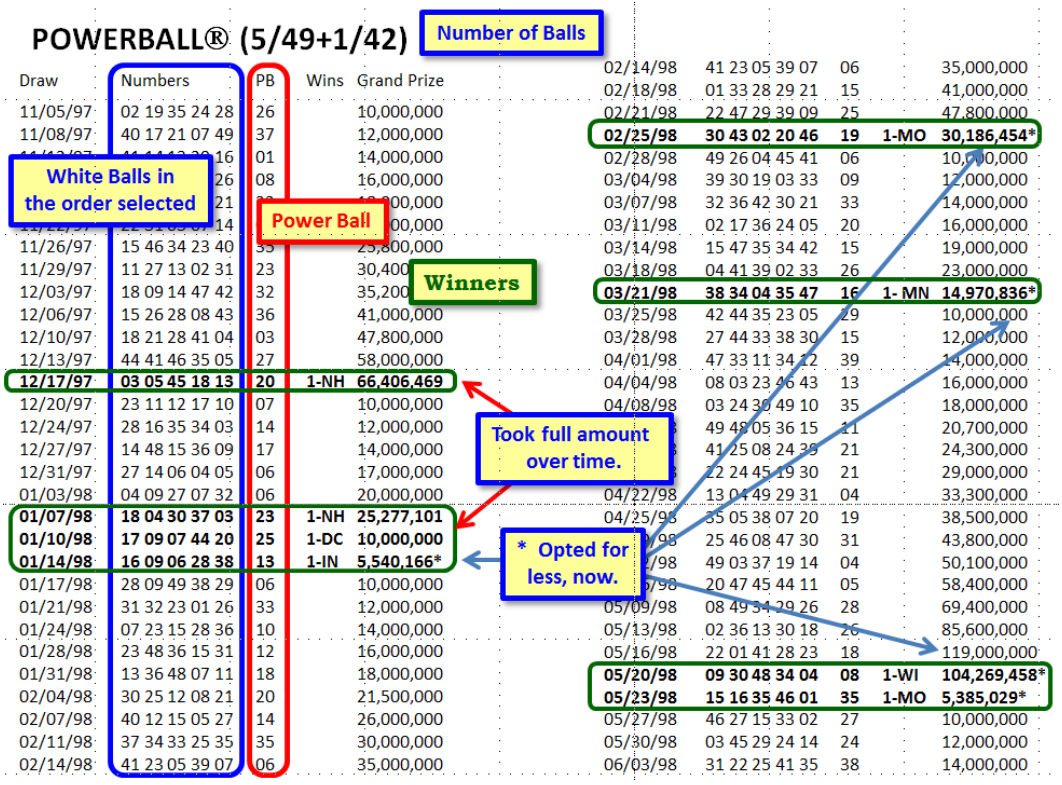
Powerball originally started in 1988 as Lotto America. The lottery changed its name to Powerball in April, 1992 with jackpots starting at $2 Million. Since then, the lottery jackpot was hit over 150 times by over 176 winners with paid out jackpot prizes (both annuity and lump sum cash) exceeding $5 billion. It is played in 42 States, Washington D.C. and the US Virgin Islands. The number of balls used changes over time.

Supporters of Powerball use a wide range of justifications. "A lottery will help fund education." "Why should we be giving our potential tax revenue to other states?" "We need to use a lottery to raise money for education so that our children will be become smart enough to understand that a lottery is really a nonproductive method for raising money." "Somebody has to win! It might as well be me."

3. The Data

Several web sites provide the Powerball numbers; the data from the site illustrated in Figure 1 gives the most useful information. After going to the web site, click on the small *hot spot* under the three most recent drawing results that provides data in Microsoft Word format. Data for lottery drawings beginning November 5, 1997 to the present are given. The first column contains the drawing dates, Wednesdays and Saturdays. The "white ball" numbers appear in the order in which they were drawn, followed by the "red ball" number. When a jackpot is won, the number of winners and the states in which the tickets were sold are given. The amount of each jackpot is last. If there was a winner,

Figure 1. Data from: **http://www.powerball.com/powerball/pb_nbr_history.asp**



Draw	Numbers	PB	Wins	Grand Prize
11/05/97	02 19 35 24 28	26		10,000,000
11/08/97	40 17 21 07 49	37		12,000,000
[illegible]	[illegible] 16	01		14,000,000
[illegible]	[illegible] 26	08		16,000,000
[illegible]	[illegible] 21	[illegible]		[illegible],000,000
[illegible]	[illegible] 14	[illegible]		[illegible],000,000
11/26/97	15 46 34 23 40	35		25,800,000
11/29/97	11 27 13 02 31	23		30,400,000
12/03/97	18 09 14 47 42	32		35,200,000
12/06/97	15 26 28 08 43	36		41,000,000
12/10/97	18 21 28 41 04	03		47,800,000
12/13/97	44 41 46 35 05	27		58,000,000
12/17/97	**03 05 45 18 13**	**20**	**1-NH**	**66,406,469**
12/20/97	23 11 12 17 10	07		10,000,000
12/24/97	28 16 35 34 03	14		12,000,000
12/27/97	14 48 15 36 09	17		14,000,000
12/31/97	27 14 06 04 05	06		17,000,000
01/03/98	04 09 27 07 32	06		20,000,000
01/07/98	**18 04 30 37 03**	**23**	**1-NH**	**25,277,101**
01/10/98	**17 09 07 44 20**	**25**	**1-DC**	**10,000,000**
01/14/98	**16 09 06 28 38**	**13**	**1-IN**	**5,540,166***
01/17/98	28 09 49 38 29	06		10,000,000
01/21/98	31 32 23 01 26	33		12,000,000
01/24/98	07 23 15 28 36	10		14,000,000
01/28/98	23 48 36 15 31	12		16,000,000
01/31/98	13 36 48 07 11	18		18,000,000
02/04/98	30 25 12 08 21	20		21,500,000
02/07/98	40 12 15 05 27	14		26,000,000
02/11/98	37 34 33 25 35	35		30,000,000
02/14/98	41 23 05 39 07	06		35,000,000
02/18/98	01 33 28 29 21	15		41,000,000
02/21/98	22 47 29 39 09	25		47,800,000
02/25/98	**30 43 02 20 46**	**19**	**1-MO**	**30,186,454***
02/28/98	49 26 04 45 41	06		10,000,000
03/04/98	39 30 19 03 33	09		12,000,000
03/07/98	32 36 42 30 21	33		14,000,000
03/11/98	02 17 36 24 05	20		16,000,000
03/14/98	15 47 35 34 42	15		19,000,000
03/18/98	04 41 39 02 33	26		23,000,000
03/21/98	**38 34 04 35 47**	**16**	**1- MN**	**14,970,836***
03/25/98	42 44 35 23 05	29		10,000,000
03/28/98	27 44 33 38 30	15		12,000,000
04/01/98	47 33 11 34 12	39		14,000,000
04/04/98	08 03 23 46 43	13		16,000,000
04/08/98	03 24 30 49 10	35		18,000,000
[illegible]	49 48 05 36 15	[illegible]		20,700,000
[illegible]	41 25 08 24 39	21		24,300,000
[illegible]	22 24 45 19 30	21		29,000,000
04/22/98	13 04 49 29 31	04		33,300,000
04/25/98	35 05 38 07 20	19		38,500,000
[illegible]	25 46 08 47 30	31		43,800,000
[illegible]	49 03 37 19 14	04		50,100,000
[illegible]	20 47 45 44 11	05		58,400,000
05/09/98	08 49 34 29 26	28		69,400,000
05/13/98	02 36 13 30 18	26		85,600,000
05/16/98	22 01 41 28 23	18		119,000,000
05/20/98	**09 30 48 34 04**	**08**	**1-WI**	**104,269,458***
05/23/98	**15 16 35 46 01**	**35**	**1-MO**	**5,385,029***
05/27/98	46 27 15 33 02	27		10,000,000
05/30/98	03 45 29 24 14	24		12,000,000
06/03/98	31 22 25 41 35	38		14,000,000

the line is **bold**. If the winner opted to accept a lesser amount now, an asterisk appears after the amount. Otherwise, the winner chose the full amount to be paid out over time.

You can find these data in a more user friendly format at the website given in Section 1. Four separate files are provided.

Data Set 1. Data from the beginning 04/22/92 to 11/01/97. It contains only the numbers for the five white balls sort in order from low to high and the one red power ball.

Data Set 2. Data for one thousand drawings from 11-05-9-1997 to 06-02-2007. Three sets of rules are included. White ball numbers are in the order in which they were selected. The last column has all white ball numbers stacked (n = 5,000).

Data Set 3. Data from 11/05/97 to 07/21/12. The white ball numbers are given in the order in which they were drawn and also arranged as order statistics. Additional columns are created to enable the development of a model for the amount of money in the jackpot. A table of variable definitions is included to the right of the data.

Data Set 4. A subset of Data Set 3 with 281 cases. This contains only the data needed to create a model for the exponential growth of the jackpot value. This is a more manageable data set for the modeling exercise.

4. Odds of Winning

At each drawing five white balls are selected randomly followed by one white ball. Over the years the numbers of white and red balls in the two bins have been changed to alter the odds of winning. Figure 2 gives the dates when changes were made to the game and the resulting odds of winning. The "Rules" column has the notation used in the data sets to define the rules of the game for each time period.

Figure 2. Odds of Winning based on the numbers of white and red balls.

Change Date	White	Red	Rules*	Odds to 1.	Winners
Apr 22, 1992	45	45	(5/45+1/45)	54,979,155	578
Nov 5, 1997	49	42	(5/49+1/42)	80,089,128	414
Oct 9, 2002	53	42	(5/53+1/42)	120,526,770	302
Aug 31, 2005	55	42	(5/55+1/42)	146,107,962	350
Jan 7, 2009	59	39	(5/59+1/39)	195,249,054	316
Jan 18, 2012	59	35	(5/59+1/35)	175,223,510	94

***** (5/45+1/45) = *Choose 5 of 45 white balls plus 1 of 45 red balls.*

How unusual are these odds? To put this into perspective, consider the history of major league baseball since 1900. See if you can match my pick. Select one major league team; pick one year in which they played; pick one of their home games for that year. (Teams played 77 home games through 1960, since then they played 81.) Now select one inning of that game; pick the top half or the bottom half of that inning. Finally pick one pitch from that half of the inning. In the history of major league baseball since 1900, there have been approximately 47,061,810 pitches thrown. Your likelihood of matching my pitch is 3.72 times better than winning the Powerball! Another example. Pick any

one year, month, and day in the 20th century; then pick a specific hour and minute of that day. The odds that your pick will match someone else's are 52,596,000 to 1. This is a 3.33 better chance than winning the Powerball! The history of the United States since 00:00:01 of July 5, 1776 to JSM 2012 has consisted of 124,218,720 minutes. Two people randomly picking the same minute is 1.41 times higher than winning the Powerball! Yet, winners are not rare due to the incredibly large numbers of players who buy tickets.

Student Exercise 1. Determine the numbers of white and red balls that should be used so that the odds of winning are: a) 100,000,000 to 1. b) 160,000,000 to 1.

Student Exercise 2. Are the number of winners during each time period, as given in Figure 2, consistent with the odds of winning?

Figure 3. A default histogram for Powerball data.

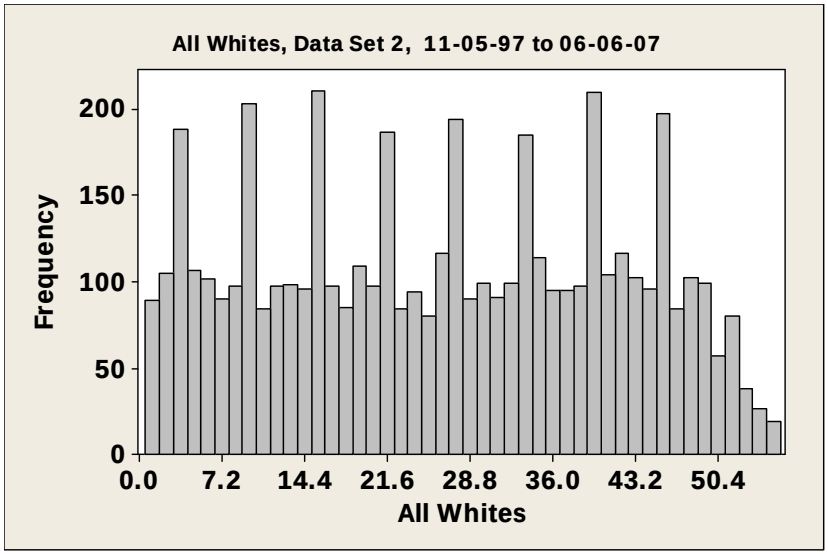
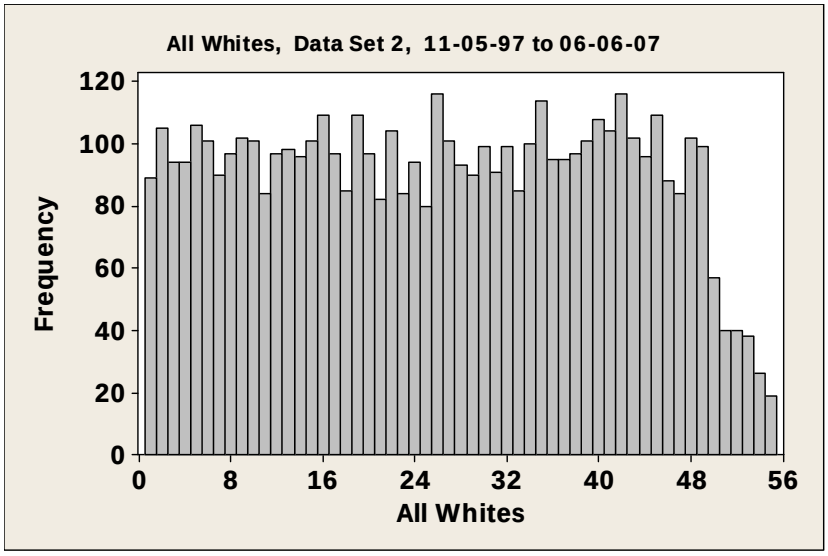


Figure 4. Histogram with one cell for each number.



5. Are some numbers "hot" or "cold"?

Default histograms can be misleading. The desirable number of bins in a histogram for continuous data ranging from 0 to 42.0 would vary depending on the distribution and the sample size. Changing the number of bins would affect the resolution. With discrete data a different numbers of possible X-values could fall into adjacent bins. Figure 3 is a default histogram having 46 bins for 55 discrete values. It gives the appearance that some values are more or less likely to occur than others. Unsuspecting students can fall into this trap and select numbers that appear to be "hot." The value of the exercise for the student is the act of discovery of the misleading information in the graph and the need to adjust. Figure 4 contains a histogram in which each discrete value has its own bin. This graph, however, presents a new dilemma. Random fluctuations appear for most of the numbers, but the frequencies for the upper bins appear to be too low. Solving this problem requires researching the history of Powerball or reviewing the data set carefully.

Assuming that the students are not provided with Figure 2, they need to determine that the rules of Powerball changed twice during the period during which these 1,000 drawings were held. The number of white balls increased from 49 to 53 to 55; hence, these numbers had less opportunity to be selected. The creative student would adjust for the differences in the time periods to see if the occurrences are indeed random.

Student Exercise 3. Students are asked to identify numbers that tend to occur more often or less often than other numbers.

6. Modeling the Growth of the Powerball Jackpot

Each time a Powerball jackpot is won, a new sequence of drawing starts. Each sequence is a time series in which the prize amount increases with each drawing. Figure 5 shows a series of these sequences. A series starts at a base amount that began at $10,000,000 in 1992 and increased to $40,000,000 in 2012. After each unsuccessful drawing the pot increases. Each black dot in Figure 5 indicates the amount of the pot for a particular drawing. A red dot indicates the reduced amount that a winner received as a buy-out for taking a lump sum rather the actual value that would be paid out over 20 years. Note the red dots are below the pot value of the previous drawing. The actual advertised value of the jack pot is considerably higher than what is indicated by the red dot amount. The red dot amount should not be used in modeling the time series. The first series shown had 10 drawings and was won on 3/27/02. The jackpot was $42,600,000 the drawing before, but the winner accepted $26,072,770. The fourth sequence was won in only three drawings; the winner took the full amount of $16,000,000 paid in an annuity. The eighth sequence grew to more than $217,800,000 before it was won on the 17^{th} drawing.

Data Set 4 is a modification of the basic data set that facilitates the ability to create a separate model for the estimating of the amount of the payout as a function of the number of drawings since the last jackpot was won. Figure 6 shows a portion of this file. Additional variables have been created to enable the sequence of drawings leading to each jack pot to be modeled individually. The first two columns give the number of winners and state in which the winning ticket was sold followed by the grand prize amount, The third column is an indicator variable: 0 if no winner, 1 if a winner. Given in Column 4 is the sequence, or pot, number, beginning on 11/05/97. "Time w/i pot" is the drawing number within each sequence; this resets to 1 each time a pot is won. This is

the X-variable used for building a model. The "Advertised Payout" has the actual jackpot value for the days the jackpot was won; this was not made available until 05/07/2011.

Figure 5. The grand prize amount versus drawing date for selected cases.

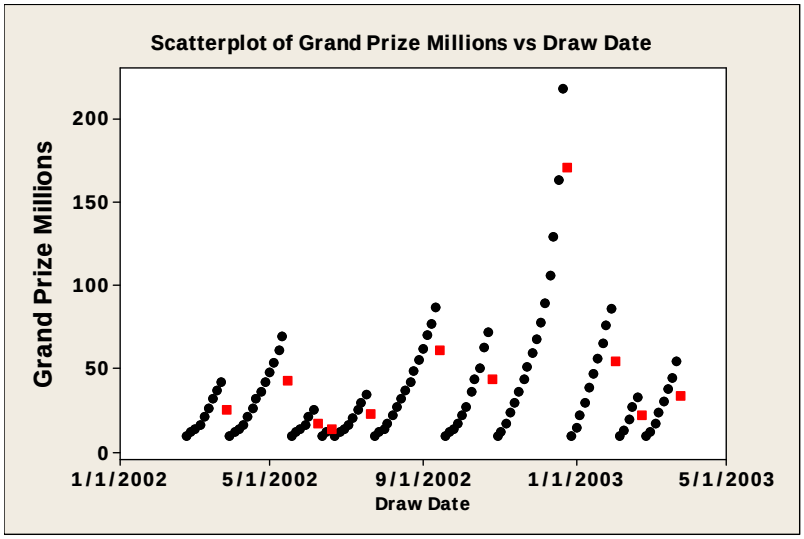


Figure 6. Layout for Data Set 4, which is a subset of Data Set 3.

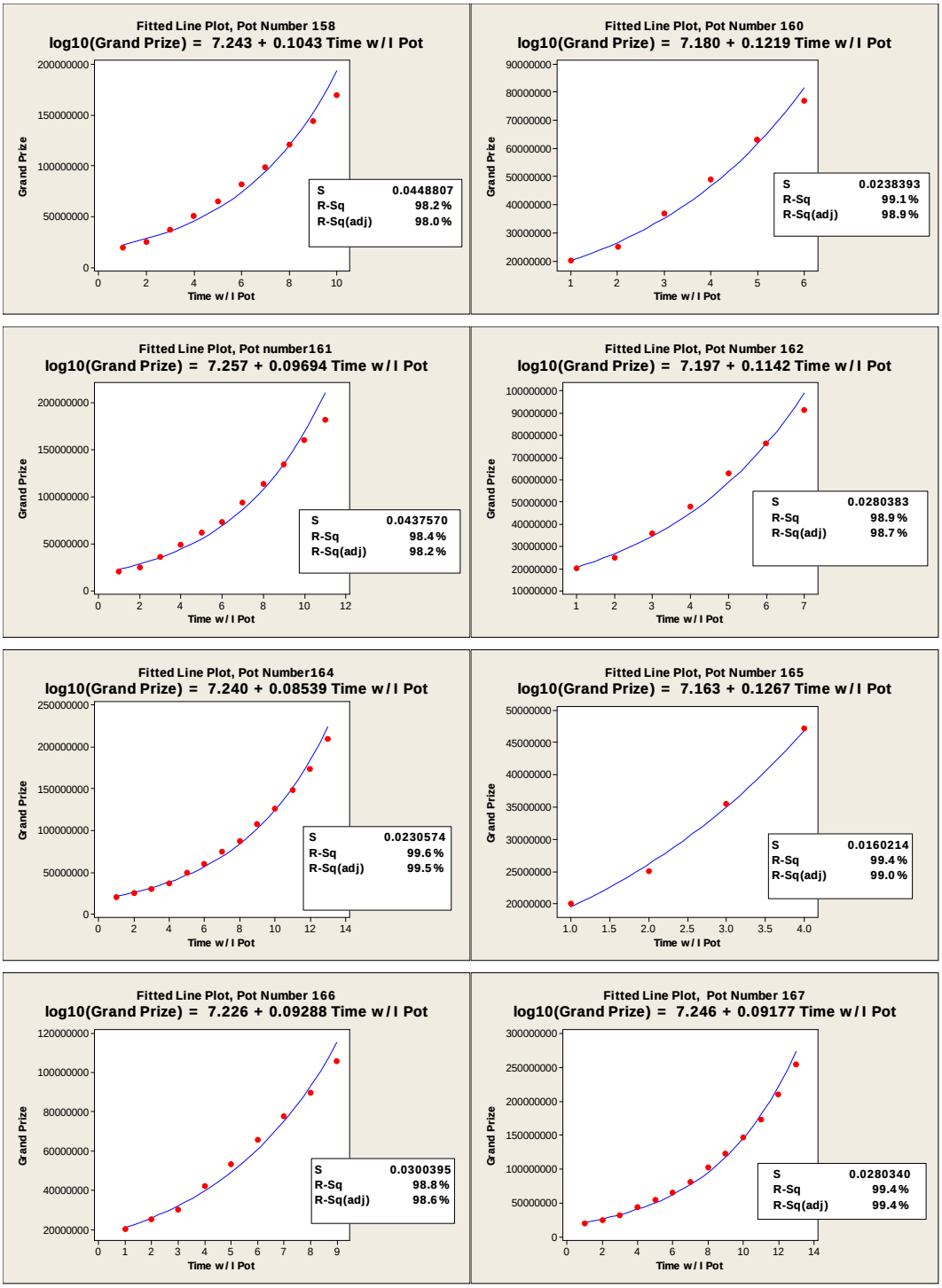
Winners	Grand Prize	If winner	Pot Num	Time w/i Pot	Game	Advertised Payout	Accepted Buyout
	20,000,000	0	158	1	5		*
	25,000,000	0	158	2	5		*
	36,900,000	0	158	3	5		*
	51,100,000	0	158	4	5	51,100,000	*
	65,200,000	0	158	5	5	65,200,000	*
	81,600,000	0	158	6	5	81,600,000	*
	98,900,000	0	158	7	5	98,900,000	*
	121,200,000	0	158	8	5	121,200,000	*
	144,200,000	0	158	9	5	144,200,000	*
	169,300,000	0	158	10	5	169,300,000	*
1-NY	*	1	158	11	5	200,000,000	106,159,637
	20,000,000	0	160	1	5	20,000,000	*
	25,000,000	0	160	2	5	25,000,000	*
	36,700,000	0	160	3	5	36,700,000	*
	49,000,000	0	160	4	5	49,000,000	*
	63,300,000	0	160	5	5	63,300,000	*
1-GA	77,100,000	0	160	6	5	77,100,000	*
	20,000,000	0	161	1	5	20,000,000	*
	25,000,000	0	161	2	5	25,000,000	*
	36,100,000	0	161	3	5	36,100,000	*
	48,600,000	0	161	4	5	48,600,000	*
	62,400,000	0	161	5	5	62,400,000	*
	73,600,000	0	161	6	5	73,600,000	*
	93,600,000	0	161	7	5	93,600,000	*
	113,600,000	0	161	8	5	113,600,000	*
	135,000,000	0	161	9	5	135,000,000	*
	160,700,000	0	161	10	5	160,700,000	*
	182,300,000	0	161	11	5	182,300,000	*
1-MN	*	1	161	12	5	220,000,000	123,602,685
	20,000,000	0	162	1	5	20,000,000	*

Regressions were run in Minitab to fit the model

$$\text{LOG10(Grand Prize)} = b_0 + b_1 \text{ Time} + e.$$

Data Set 4 was split into separate worksheets. "Fitted line plot" was used to produce the outputs shown in Figure 7. The prize amounts do appear to grow exponentially.

Figure 7. Fitted line plot outputs for regressing Log(Grand Prize) as a function of the drawing number. There is one plot for each Pot Number.



The estimated parameters for the exponential models are given in Figure 8.

Figure 8. Summary of estimated coefficients for fitting the Log(Grand Prize) to the drawing number.

Pot ID	n	b_0	b_1	r-Square
158	10	7.243	0.1043	98.2%
160	6	7.180	0.1219	99.1
161	11	7.257	0.09694	98.4
162	7	7.197	0.1142	98.9
164	13	7.240	0.08539	99.6
165	4	7.163	0.1267	99.4
166	9	7.226	0.09288	98.8
167	13	7.246	0.09177	99.4

Student Exercise 4. Students are asked to develop a model that describes the relationship between the grand prize and the drawing number. Use Data Set 4.

Student Exercise 5. Could the same model be used to model the fit for all the sequences in Data Set 4? That is, test the hypothesis that all the sequences have the same constant and coefficient for the drawing number.

Student Exercise 6. What other factors might influence the growth of the jackpots? Answers could include the time of year, recent LARGE jackpots, types of advertising used, unemployment rate, libor rate, ...

7. Discrete Order Statistics

When the numbers drawn are sorted in order of magnitude, each of the five columns contains the i^{th} order statistic. A histogram of each column shows the distribution of these order statistics. A curious student may want to know if these distributions follow any known distributions. For example, could the distribution of the first order statistic be modeled with an exponential distribution?

Figure 9 contains the histograms of the order statistics of these data from Data Set 3. Two observations are contained in each bin.

Student Exercise 7. Suppose a secondary betting market is created in which a person bets on the likelihood the lowest of the five white balls selected will be:

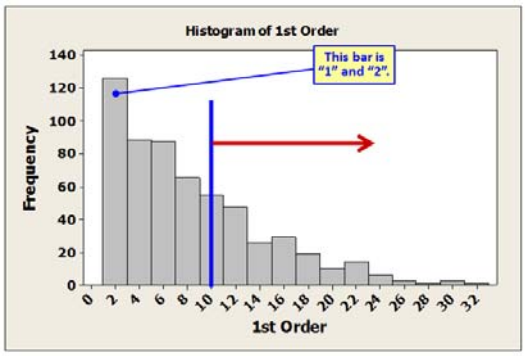
a. Greater than or equal to, say, 10. (See Figure 9a.)
b. Less than, say, 15. Estimate the probability of these events happening.
c. Find the expected value and standard deviation for each of the order statistic.

Student Exercise 8. Perform a goodness of fit:

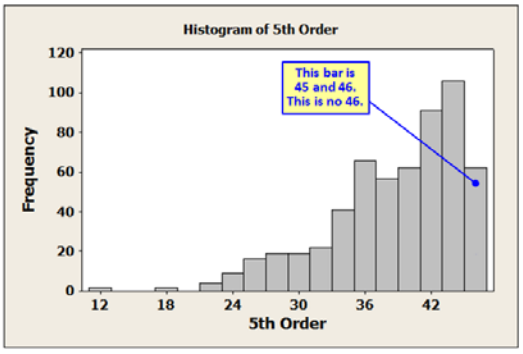
a. Does the third order statistic fit a normal distribution?
b. Does the first order statistic fit an exponential distribution?
c. What about the others?

Figure 9. Distribution of Order Statistics, Game 1: 5/45 + 1/45, 1992 - 1997

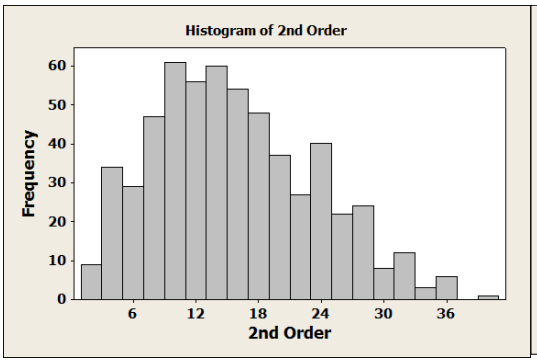
a.



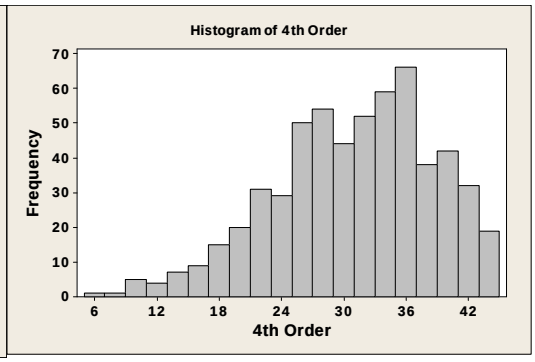
e.



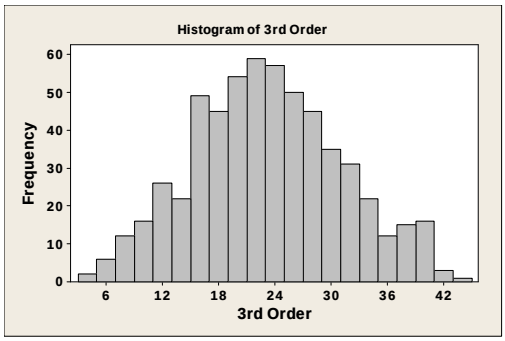
b.



d.



c.



8. Summary

Students can learn through exploring and exploiting sets of data, using imagination and skepticism, wondering why relationships among variables are or are not what they appear to be. Giving them opportunities to explore data is a benefit for the students. Powerball data is one such data set. There are many more. Some specific exercises are suggested here; however, giving fewer specifics can lead to more creativity.

References

http://www.lotteryuniverse.com/powerball-history.aspx
History of numbers in Powerball

http://answers.yahoo.com/question/index?qid=20090522120240AA4qfAr
Games per year in baseball.