Chapter 15

Descriptive and Inferential Statistics in Undergraduate Data Science Research Projects

Malcolm J. D'Souza, Edward A. Brandenburg,
Derald E. Wentzien, Riza C. Bautista,
Agashi P. Nwogbaga, Rebecca G. Miller and
Paul E. Olsen

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/65721

Abstract

Undergraduate data science research projects form an integral component of the Wesley College science and mathematics curriculum. In this chapter, we provide examples for hypothesis testing, where statistical methods or strategies are coupled with methodologies using interpolating polynomials, probability and the expected value concept in statistics. These are areas where real-world critical thinking and decision analysis applications peak a student's interest.

Keywords: Wesley College, STEM, undergraduate research, solvolysis, phenyl chloroformate, benzoyl chloride, benzoyl fluoride, benzoyl cyanide, Grunwald-Winstein equation, transition-state, addition-elimination, multiple regression, time-series, Ebola, polynomial functions, probability, expected value

1. Introduction

Wesley College (Wesley) is a minority-serving, primarily undergraduate liberal-arts institution. Its STEM (science, technology, engineering and mathematics) fields contain a robust (federal and state) sponsored directed research program [1, 2]. In this program, students receive individual mentoring on diverse projects from a full-time STEM faculty member. In addition, undergraduate research is a capstone thesis requirement and students complete research projects within experiential courses or for an annual Scholars' Day event.

Undergraduate research is indeed a hallmark of Wesley's progressive liberal-arts core-curriculum. All incoming freshmen are immersed in research in a specially designed quantitative reasoning a 100-level mathematics core course, a first-year seminar course and 100-level frontiers in science core course [1]. Projects in all level-1 STEM core courses provide an opportunity to develop a base knowledge for interacting and manipulating data. These courses also introduce students to modern computing techniques and platforms.

At the other end of the Wesley core-curriculum spectrum, the advanced undergraduate STEM research requirements reflect the breadth and rigor necessary to prepare students for (possible) future postgraduate programs. For analyzing data in experiential research projects, descriptive and inferential statistics are major components. In informatics, students are trained in the SAS Institute's statistical analysis system (SAS) software and in the use of geographic information system (GIS) spatial tools through ESRI's ArcGIS platform [2].

To help students with poor mathematical ability and to further enhance their general thinking skills, in our remedial mathematics courses, we provide a foundation in algebraic concepts, problem-solving skills, basic quantitative reasoning and simple simulations. Our institution also provides a plethora of student academic support services that include an early alert system, peer and professionally trained tutoring services and writing center support. In addition, Wesley College non-STEM majors are required to take the project-based 100-level mathematics core course and can then opt to take two project-based 300-level SAS and GIS core courses. Such students who are trained in the concepts and applications of mathematical and statistical methods can then participate in Scholars' Day to augment their mathematical and critical thinking skills.

2. Linear free energy relationships to understand molecular pathways

Single and multiparameter linear free energy relationships (LFERs) help chemists evaluate multiple kinds of transition-state molecular interactions observed in association with compound variability [3]. Chemical kinetics measurements are understood by correlating the experimental compound reaction rate (*k*) or equilibrium data and their thermodynamics. The computationally challenging stoichiometric analysis elucidates metabolic pathways by analyzing the effect of physiochemical, environmental and biological factors on the overall chemical network structure. All of these determinations are important in the design of chemical processes for petrochemical, pharmaceutical and agricultural building blocks.

In this section, through results obtained from our undergraduate directed research program in chemistry, we outline examples with statistical descriptors that use inferential correctness for testing hypotheses about regression coefficients in LFERs that are common to the study of solvent reactions. To understand mechanistic approaches, multiple regression correlation analyses using the one- and two-term Grunwald-Winstein equations (Eqs. (1) and (2)) are proven to be effective instruments that elucidate the transition-state in solvolytic reactions [3]. To avoid multicollinearity, it is stressed that the chosen solvents have widely varying ranges of nucleophilicity (*N*) and solvent-ionizing power (*Y*) values [3, 4]. In Eqs. (1) and (2) (for a particular substrate), k is the rate of reaction in a given solvent, k_o is the 80% ethanol (EtOH) reaction rate, l is the sensitivity toward changes in N, m is the sensitivity toward changes in Y and c is a constant (residual) term. In substrates that have the potential for transition-state electron delocalization, Kevill and D'Souza introduced an additional hI term to Eqs. (1) and (2) (and as shown in Eqs. (3) and (4)). In Eqs. (3) and (4), h represents the sensitivity to changes in the aromatic ring parameter I [3].

$$\log\left(k/k_o\right) = mY + c \tag{1}$$

$$\log\left(k/k_o\right) = lN + mY + c \tag{2}$$

$$\log\left(k/k_o\right) = mY + h\,I + c \tag{3}$$

$$\log\left(k/k_o\right) = lN + mY + hI + c \tag{4}$$

Eqs. (1) and (3) are useful in substrates where the unimolecular dissociative transition-state (S_N1 or E1) formation is rate-determining. Eqs. (2) and (4) are employed for reactions where there is evidence for bimolecular associative (S_N2 or E2) mechanisms or addition-elimination (A-E) processes. In substrates undergoing similar mechanisms, the resultant l/m ratios obtained can be important indicators to compensate for earlier and later transition-states (TS). Furthermore, l/m ratios between 0.5 and 1.0 are indicative of unimolecular processes (S_N1 or E1), values ≥ 2.0 are typical in bimolecular processes (S_N2, E2, or A-E mechanisms) and values <<0.5 imply that ionization-fragmentation is occurring [3].

To study the (solvent) nucleophilic attack at a sp^2 carbonyl carbon, we completed detailed Grunwald-Winstein (Eqs. (1), (2) and (4)) analyses for phenyl chloroformate (PhOCOCl) at 25.0°C in 49 solvents with widely varying N and Y values [3, 4]. Using Eq. (1), we obtained an m value of -0.07 ± 0.11, $c = -0.46 \pm 0.31$, a very poor correlation coefficient ($R = 0.093$) and an extremely low F-test value of 0.4. An analysis of Eq. (2) resulted in a very robust correlation, with $R = 0.980$, F-test = 568, $l = 1.66 \pm 0.05$, $m = 0.56 \pm 0.03$ and $c = 0.15 \pm 0.07$. Using Eq. (4), we obtained $l = 1.77 \pm 0.08$, $m = 0.61 \pm 0.04$, $h = 0.35 \pm 0.19$ (P-value = 0.07), $c = 0.16 \pm 0.06$, $R = 0.982$ and the F-test value was 400.

Since the use of Eq. (2) provided superior statistically significant results (R, F-test and P-values) for PhOCOCl, we strongly recommended that in substrates where nucleophilic attack occurs at a sp^2 hybridized carbonyl carbon, the PhOCOCl l/m ratio of 2.96 should be used as a guiding indicator for determining the presence of an addition-elimination (A-E) process [3, 4]. Furthermore, for n-octyl fluoroformate (OctOCOF) and n-octyl chloroformate (OctOCOCl), we found that the leaving-group ratio (k_F/k_{Cl}) was close to, or above unity. Fluorine is a very poor leaving-group when compared to chlorine, hence for carbonyl group containing molecules, we proposed the existence of a bimolecular tetrahedral transition-state (TS) with a rate-determining addition step within an A-E pathway (as opposed to a bimolecular concerted associative S_N2 process with a penta-coordinate TS).

For chemoselectivity, the sp^2 hybridized benzoyl groups (PhCO—) are found to be efficient and practical protecting agents that are utilized during the synthesis of nucleoside, nucleotide and oligonucleotide analogue derivative compounds. Yields for regio- and stereoselective reactions are shown to depend on the preference of the leaving group and commercially, benzoyl fluoride (PhCOF), benzoyl chloride (PhCOCl) and benzoyl cyanide (PhCOCN) are cheap and readily available.

We experimentally measured the solvolytic rates for PhCOF at 25.0°C [5]. In 37 solvent systems, a two-term Grunwald-Winstein (Eq. (2)) application resulted in an l value of 1.58 ± 0.09, an m value of 0.82 ± 0.05, a c value of −0.09, R = 0.953 and the F-test value was 186. The l/m ratio of 1.93 for PhCOF is close to the OctOCOF l/m ratio of 2.28 (in 28 pure and binary mixtures) indicating similar A-E transition states with rate-determining addition.

On the other hand, for PhCOCl at 25.0°C, we used the available literature data (47 solvents) from various international groups and proved the presence of simultaneous competing dual side-by-side mechanisms [6]. For 32 of the more ionizing solvents, we obtained $l = 0.47 \pm 0.03$, $m = 0.79 \pm 0.02$, $c = -0.49 \pm 0.17$, $R = 0.990$ and F-test = 680. The l/m ratio is 0.59. Hence, we proposed an S_N1 process with significant solvation (l component) of the developing aryl acylium ion. In 12 of the more nucleophilic solvents, we obtained $l = 1.27 \pm 0.29$, $m = 0.46 \pm 0.07$, $c = 0.18 \pm 0.23$, $R = 0.917$ and F-test = 24. The l/m ratio of 2.76 is close to the 2.96 value obtained for PhOCOCl. This suggests that the A-E pathway is prevalent. In addition, there were three solvents where there was no clear demarcation of the changeover region.

At 25.0°C in solvents that are common to PhCOCl and PhCOCF we observed $k_{\mathrm{PhCOCl}} > k_{\mathrm{PhCOF}}$. This rate trend is primarily due to more efficient PhCOF ground-state stabilization.

Lee and co-workers followed the kinetics of benzoyl cyanide (PhCOCN) at 1, 5, 10, 15 and 20°C in a variety of pure and mixed solvents and proposed the presence of an associative S_N2 (penta-coordinate TS) process [7]. PhCOCN is an ecologically important chemical defensive secretion of polydesmoid millipedes and cyanide is a synthetically useful highly active leaving group. Since the leaving group is involved in the rate-determining step of any S_N2 process, we became skeptical with the associative S_N2 proposal and decided to reinvestigate the PhCOCN analysis. We hypothesized that since PhCOCl showed mechanism duality, similar analogous dual mechanisms should endure during PhCOCN solvolyses.

Using the Lee data within Arrhenius plots (Eq. (5)), we determined the PhCOCN solvolytic rates at 25°C (**Table 1**). We obtained the rates for PhCOCN in 39 pure and mixed

$$\ln(k) = \frac{-Ea}{RT} + \ln(A) \tag{5}$$

Solvent (v/v)	10^5 k/s^{-1}	N_T	Y_{cl}	I	Solvent (v/v)	10^5 k/s^{-1}	N_T	Y_{cl}	I
90% EtOH[1]	139.9	0.16	−0.94	0.10	**30% acetone**[1]	2447	−0.96	3.21	−0.38
80% EtOH[1]	210.0	0.00	0.00	0.00	**20% acetone**[1]	3726	−1.11	3.77	−0.40
70% EtOH[1]	322.8	−0.20	0.78	−0.06	**10% acetone**[1]	4071	−1.23	4.28	−0.43[4]
60% EtOH[1]	598.8	−0.38	1.38	−0.15	**30% dioxane**[1]	1690	−0.98	2.97	−0.29[5]
50% EtOH[1]	986.7	−0.58	2.02	−0.23	**20% dioxane**[1]	2887	−1.12	3.71	−0.25
40% EtOH[1]	1761	−0.74	2.75	−0.24	**10% dioxane**[1]	4196	−1.25	4.23	−0.34[5]
30% EtOH[1]	3064	−0.93	3.53	−0.30	**10T–90E**[2]	37.20	0.27[4]	−1.99[4]	0.26[4]
20% EtOH[1]	3732	−1.16	4.09	−0.33	**20T–80E**[2]	41.64	0.08	−1.42	0.31
90% MeOH[1]	390.5	−0.01	−0.18	0.28	**30T–70E**[2]	35.50	−0.11	−0.95	0.38
80% MeOH[1]	575.3	−0.06	0.67	0.14	**40T–60E**[2]	32.98	−0.34	−0.48	0.43
70% MeOH[1]	800.2	−0.40	1.46	0.04	**50T–50E**[2]	30.30	−0.64	0.16	0.51
60% MeOH[1]	1616	−0.54	2.07	−0.19	**70T–30E**[2]	27.70	−1.34	1.24	0.65[4]
50% MeOH[1]	2573	−0.75	2.70	−0.05	**76.3 TFE**[3]	639.9	−2.19[4]	2.84	0.28
40% MeOH[2]	4205	−0.87	3.25	−0.13	**67.4 TFE**[3]	886.1	−1.88[4]	2.93[4]	0.22[4]
30% MeOH[2]	5351	−1.06	3.73	−0.22	**57.9 TFE**[3]	1075	−1.78[4]	3.05[4]	0.14[4]
80% acetone[1]	9.547	−0.37	−0.83	−0.23	**47.9 TFE**[3]	1512	−1.33[4]	3.21[4]	0.06[4]
70% acetone[1]	86.02	−0.42	0.17	−0.29	**37.1 TFE**[3]	2089	−1.19[4]	3.44[4]	−0.03[4]
60% acetone[1]	157.1	−0.52	0.95	−0.28	**25.6 TFE**[3]	2944	−1.15[4]	3.73[4]	−0.15[4]
50% acetone[1]	505.9	−0.70	1.73	−0.32	**13.3 TFE**[3]	3870	−1.23[4]	4.10[4]	−0.29[4]
40% acetone[1]	1149	−0.83	2.46	−0.35	–	–	–	–	–

[1] Calculated using four data points in an Arrhenius plot.
[2] Calculated using three data points in an Arrhenius plot.
[3] Calculated using three data points in an Arrhenius plot and are w/w compositions.
[4] Determined using a second-degree polynomial equation.
[5] Determined using a third-degree polynomial equation.

Table 1. The 25.0°C calculated rates for PhCOCN, the N_T, Y_{Cl} and I values.

aqueous organic solvents of ethanol (EtOH), methanol (MeOH), acetone (Me_2CO), dioxane, 2,2,2-trifluoroethanol (TFE) and in TFE-EtOH (T-E) mixtures. For all of the Arrhenius plots, the R^2 values ranged from 0.9937 to 1.0000, except in 60% Me_2CO, R^2 was 0.9861. The Arrhenius plot for 80% EtOH is shown in **Figure 1**. In order to utilize Eqs. (1)–(4) for all 39 solvents, second degree or third-degree polynomial equations were used to calculate the missing N_T, Y_{Cl} and I values. The calculated 25°C PhCOCN reaction rates and the literature available or interpolated N_T, Y_{Cl} and I values are listed in **Table 1**.

Using Eq. (2) for 32 of the PhCOCN solvents in **Table 1** (20–90% EtOH, 30–90% MeOH, 20–80% Me_2CO, 10–30% dioxane, 10T–90E, 20T–80E, 30T–70E, 40T–60E, 50T–50E and 70T–30E), we obtained R = 0.988, F-test = 595, l = 1.54 ± 0.11, m = 0.74 ± 0.03 and c = 0.13 ± 0.04. Using Eq. (4), we obtained R = 0.989, F-test = 432, l = 1.62 ± 0.11, m = 0.78 ± 0.03, h = 0.22 ± 0.11 (P-value = 0.07) and c = 0.13 ± 0.04.

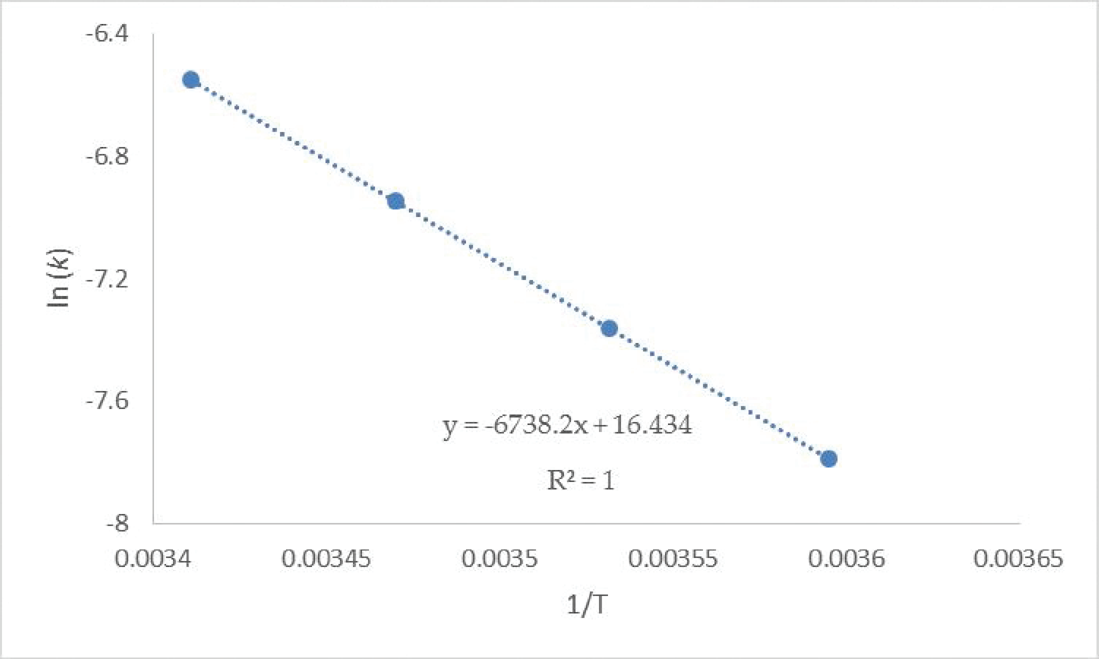

Figure 1. Arrhenius plot for 80% EtOH.

The l/m ratio of 2.08 obtained (for PhCOCN) using Eq. (2) is close to that obtained (1.93) for PhCOF and hence we propose a parallel A-E mechanism.

For the seven highly ionizing aqueous TFE mixtures, using Eq. (1) we obtained, R = 0.977, F-test = 105, m = 0.61 ± 0.06 and c = −1.15 ± 0.20. Using Eq. (2) we obtained R = 0.999, F-test = 763, l = 0.25 ± 0.031, m = 0.42 ± 0.03 and c = −0.13 ± 0.14. Using Eqs. (3) and (4) we obtained R = 0.998, F-test = 417, m = −0.65 ± 0.22 (P-value = 0.04), h = −2.83 ± 0.491 (P-value = 0.01) and c = 3.12 ± 0.73 (P-value = 0.01) and R = 0.989, F-test = 572, l = 0.17 ± 0.07 (P-value = 0.11), m = 0.02 ± 0.33 (P-value = 0.96), h = −1.04 ± 0.86 (P-value = 0.31), c = 1.10 ± 1.02 (P-value = 0.36), respectively.

In the very polar TFE mixtures, in Eq. (2) the l/m ratio was 0.60, indicating a dissociative S_N1 process. The l value of 0.25 is consistent with the need of small preferential solvation to stabilize the developing S_N1 carbocation and the lower m value (0.42) attained can be rationalized in terms of less demand for solvation of the cyanide anion (leaving group).

In all of the common solvents at 25.0°C, $k_{PhCOCl} > k_{PhCOCN} > k_{PhCOF}$. In addition, PhCOCN was found to be faster than PhCOF by a factor of 18–71 times in the aqueous ethanol, methanol, acetone and dioxane mixtures and 185–1100 times faster in the TFE-EtOH and TFE-H_2O mixtures. These observations are very reasonable as the cyanide group is shown to have a greater inductive effect and in addition, the cyanide anion is a weak conjugate base. This rationalization is logical as $(l/m)_{PhCOCN} > (l/m)_{PhCOF}$.

3. Estimating missing values from a time series data set

Complete historical data time series are needed to create effective mathematical models. Unfortunately, systems that track and record the data values periodically malfunction thereby creating missing and/or inaccurate values in the time series. If a reasonable estimate for the missing value can be determined, the data series can then be used for future analysis.

In this section, we present a methodology to generate a reasonable estimate for a missing or inaccurate values when two important conditions exist: (1) a similar data series with complete information is available and (2) a pattern (or trend) is observable.

The extent of the ice at the northern polar ice cap in square kilometers is tracked on a daily basis and this data is made available to researchers by the National Snow & Ice Data Center (NSIDC). A review of the NASA Distributed Active Archive Center (DAAC) data at NSIDC indicates that the extent of the northern polar ice cap follows a cyclical pattern throughout the year. The extent increases until it reaches a maximum for the year in mid-March and decreases until it reaches a minimum for the year in mid-September. Unfortunately, the data set contains missing data for some of the days.

The extent of the northern polar ice cap in the month of January for 2011, 2012 and 2013 is utilized as an example. Complete daily data for January in 2011 and 2012 is available. The 2013 January data has a missing data value for January 25, 2013.

Figure 2 presents the line graph of the daily ice extent for January of 2011, 2012 and 2013. A complete time series is available for 2011 and 2012, so the first condition is met. The line graphs also indicate that the extent of the polar ice caps is increasing in January, so the second condition is met. An interpolating polynomial will be introduced and used to estimate the missing value for the extent of the polar ice cap on January 25, 2013.

Let t = the time period or observation number in a time series.

Let $f(t)$ = the extent of the sea ice for time period t.

The extent of the sea ice can be written as a function of time.

For a polynomial of degree 1, the function will be: $f(t) = a_0 + a_1(t)$

For a polynomial of degree 3, the function will be: $f(t) = a_0 + a_1(t) + a_2(t)^2 + a_3(t)^3$

Polynomials of higher degrees could also be used. The extent of the polar ice for January 25 will be removed from the data series for 2011 and 2012 and an estimate will be prepared using polynomials of degree 1. Another estimate is prepared using polynomials of degree 3. The estimated value will be compared to the actual value for the years 2011 and 2012. The degree of the polynomial that generates the best (closest) estimate for January 25 will be the degree of the polynomial used to generate the estimate for January 25, 2013.

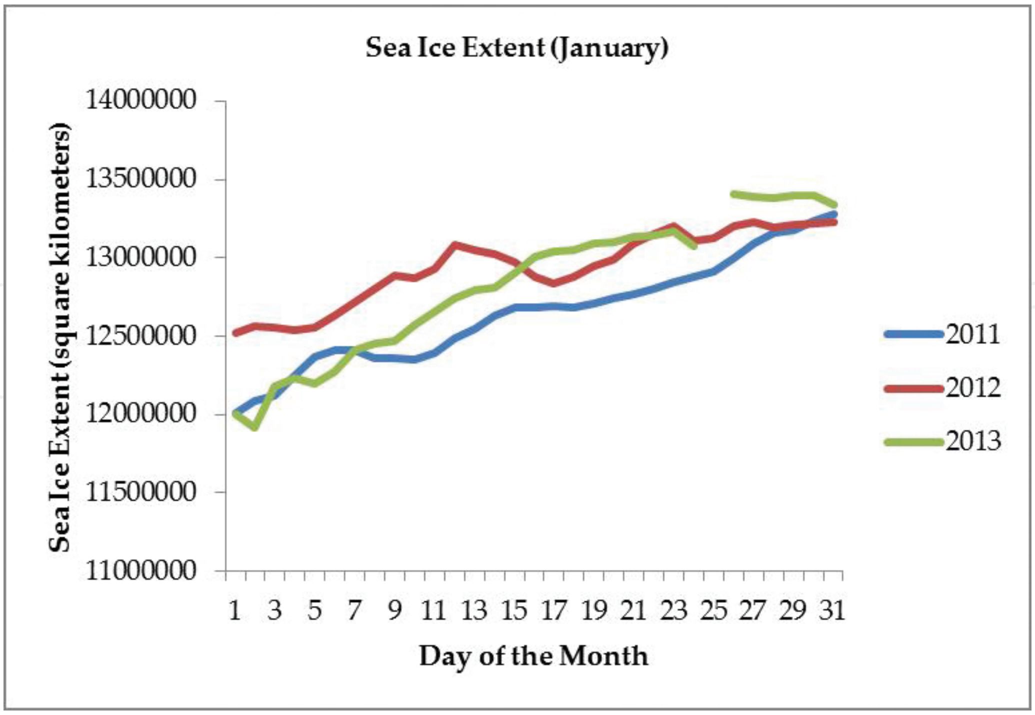

Figure 2. The extent of sea ice in January 2011, 2012 and 2013.

A two-equation, two-unknown system of equations is created when using polynomials of degree 1. One known value before and after the missing value for each year is used to set up the system of equations. To simplify the calculations, January 24 is recorded as time period 1, January 25 is recorded as time period 2 and January 26 is recorded as time period 3. The time period and extent of the sea ice for each year was recorded in Excel.

Time period	2011	2012	2013
1	12,878,750	13,110,000	13,077,813
2	12,916,563	13,123,125	
3	12,996,875	13,204,219	13,404,688

The system of equations using a first-order polynomial for January 2011 is:

$$\begin{aligned} a_0 + a_1(1) &= 12{,}878{,}750 \\ a_0 + a_1(3) &= 12{,}996{,}875 \end{aligned} \tag{6}$$

The coefficients a_i can be found by solving the system of equations. Substitution, elimination, or matrices can be used to solve the system of equations. A TI-84 graphing calculator and matrices were used to solve this system.

The solution to this system of equations is: $a_0 = 12,819,687.5, a_1 = 59,062.5$

The estimate for January 25, 2011 is: $12,819,687.5 + 59,062.2(2) = 12,937,812.5\ \mathrm{km}^2$.

The system of equations using a first-order polynomial for 2012 is:

$$\begin{aligned} a_0 + a_1(1) &= 13{,}110{,}000 \\ a_0 + a_1(3) &= 13{,}204{,}219 \end{aligned} \tag{7}$$

The solution to this system of equations is: $a_0 = 13,062,890.5, a_1 = 47,109.5$

The estimate for January 25, 2012 is: $13,062,890.5 + 47,109.5(2) = 13,157,109.5\mathrm{km}^2$.

The absolute values of the deviations (actual and estimated values) were calculated in Excel.

Degree	Year	Actual	Estimated	Absolute deviation
1	2011	12,916,563	12,937,812.5	21,249.5
1	2012	13,123,125	13,157,109.5	33,984.5

A four-equation, four-unknown system of equations is created when using polynomials of degree 3. Two known values before and after the missing value are used to set up the system of equations. To simplify the calculations, January 23 is recorded as time period 1, January 24 is recorded as time period 2, January 25 is recorded as time period 3, January 26 is recorded as time period 4 and January 27 is recorded as time period 5. The time period and extent of the sea ice for each year was recorded in Excel.

Time period	2011	2012	2013
1	12,848,281	13,199,375	13,168,594
2	12,878,750	13,110,000	13,077,813
3	12,916,563	13,123,125	
4	12,996,875	13,204,219	13,404,688
5	13,090,625	13,227,344	13,388,750

The system of equations using a third-order polynomial for 2011 is:

$$\begin{aligned} a_0 + a_1(1) + a_2(1)^2 + a_3(1)^3 &= 12{,}848{,}281 \\ a_0 + a_1(2) + a_2(2)^2 + a_3(2)^3 &= 12{,}878{,}750 \\ a_0 + a_1(4) + a_2(4)^2 + a_3(4)^3 &= 12{,}996{,}875 \\ a_0 + a_1(5) + a_2(5)^2 + a_3(5)^3 &= 13{,}090{,}625 \end{aligned} \tag{8}$$

The solution to this system of equations is: $a_0 = 12,832,811.67, a_1 = 8,985.17$, $a_2 = 5,976.33, a_3 = 507.83$

The estimate for January 25, 2011 is: $12,832,811.67 + 8,985.17(3) + 5,976.33(3)^2 + 507.33(3)^3 = 12,927,252.1\text{km}^2$.

The system of equations using a third-order polynomial for 2012 is:

$$\begin{aligned} a_0 + a_1(1) + a_2(1)^2 + a_3(1)^3 &= 13{,}199{,}375 \\ a_0 + a_1(2) + a_2(2)^2 + a_3(2)^3 &= 13{,}110{,}000 \\ a_0 + a_1(4) + a_2(4)^2 + a_3(4)^3 &= 13{,}204{,}219 \\ a_0 + a_1(5) + a_2(5)^2 + a_3(5)^3 &= 13{,}227{,}344 \end{aligned} \tag{9}$$

The solution to this system of equations is: $a_0 = 13,486,719$, $a_1 = -413,073.33$, $a_2 = 139,101.75$, $a_3 = -13,372.42$

The estimate for January 25, 2012 is: $13,486,719 - 413,073.33(3) + 139,101.75(3)^2 - 13,372.42(3)^3 = 13,138,359.42\ \ \text{km}^2$

The absolute values of the deviations (actual and estimated values) were calculated in Excel.

Degree	Year	Actual	Estimated	Absolute deviation
3	2011	12,916,563	12,927,252.1	10,689.1
3	2012	13,123,125	13,138,359.4	15,234.4

The mean of the absolute deviations for polynomials of degree 1 and the mean of the absolute deviations for polynomials of degree 3 were calculated in Excel. The polynomial of degree 3 provided the smallest mean absolute deviation.

Degree	Mean absolute deviation
1	27,617.00
3	12,961.75

Therefore, a third order polynomial will be used to generate an estimate for the sea ice extent on January 25, 2013.

The system of equations using a third-order polynomial for 2013 is:

$$\begin{aligned} a_0 + a_1(1) + a_2(1)^2 + a_3(1)^3 &= 13{,}168{,}594 \\ a_0 + a_1(2) + a_2(2)^2 + a_3(2)^3 &= 13{,}077{,}813 \\ a_0 + a_1(4) + a_2(4)^2 + a_3(4)^3 &= 13{,}404{,}688 \\ a_0 + a_1(5) + a_2(5)^2 + a_3(5)^3 &= 13{,}388{,}750 \end{aligned} \tag{10}$$

The solution to this system of equations is: $a_0 = 13,717,916.67$, $a_1 = -850,859.17$, $a_2 = 337,669.33$, $a_3 = -36,132.83$.

The estimate for January 25, 2013 is: $13,717,916.67 - 850,859.17(3) + 337,669.33(3)^2 - 36,132.83(3)^3 = 13,228,776.72$ km^2. **Figure 3** shows the extent of the sea ice in January, 2013 with the estimate for January 25.

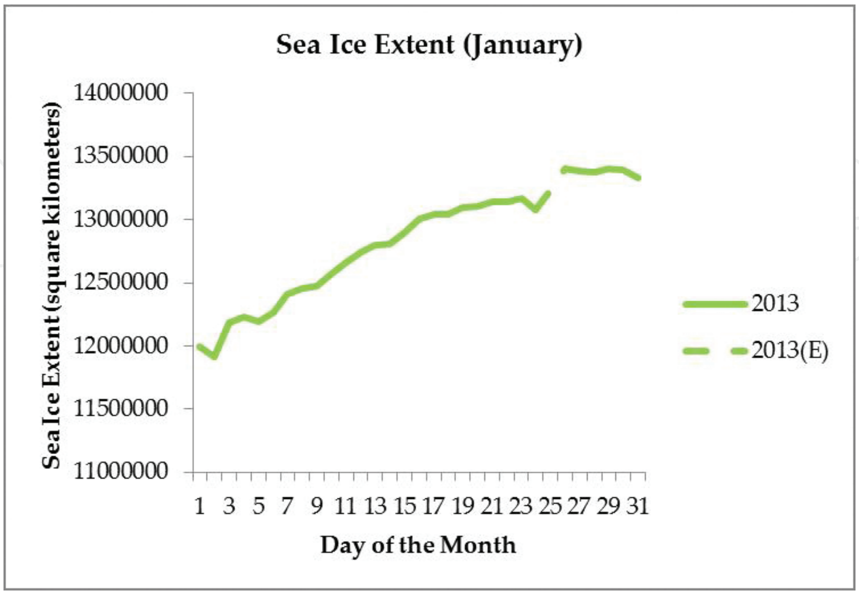

Figure 3. The extent of sea ice in January 2013 with the January 25, 2013 estimate.

4. Statistical methodologies and applications in the Ebola war

In 2014, an unprecedented outbreak of Ebola occurred predominantly in West Africa. According to the Center for Disease Control (CDC), over 28.5 thousand cases were reported resulting in more than 11,000 deaths [8]. The countries that were affected by the Ebola outbreak were Senegal, Guinea, Nigeria, Mali, Sierra Leone, Liberia, Spain and the United States of America (USA). Statistics through dynamic modeling played a crucial role with clinical data collection and management. The lessons learned and the resultant statistical advances continue to inform and drive current and subsequent pandemics.

For this honors thesis project, we tracked and gathered Ebola data over an extended period of time from the CDC, World Health Organization (WHO) and the news media [8, 9]. We used statistical curve fitting that involved both exponential and polynomial functions as well as model validation using nonlinear regression and R^2 statistical analysis.

The first WHO report (initial announcement) of the West Africa Ebola outbreak was made during the March 23rd, 2014 week. Consequently, the data for this project began from that week to October 31, 2014. The 2014 Ebola data was used to create epidemiological models to predict the possible pathway of a 2014 West Africa type of Ebola outbreak. The WHO number of Ebola cases and death toll as of October 31st, 2014 were Liberia (6635 cases with 2413 deaths), Sierra Leone (5338 cases with 1510 deaths), Guinea (1667 cases with 1018 deaths), Nigeria (20 cases with eight deaths), the United States (four cases with one death), Mali (one case with one death) and Spain (one case with zero death).

Microsoft Excel was used for the modeling of the three examples shown and were predicated upon the following assumptions: (1) Week 1 is the week of March 23rd, 2014; (2) X is the number of weeks starting from Week 1 and Y is the number of Ebola deaths; (3) there was no vaccine/cure; and (4) the missing data for the 24th week was obtained by interpolation.

4.1. Modeling of weekly Guinea Ebola deaths

The dotted curve in **Figure 4** shows the actual observed deaths while the solid line shows the number of deaths as determined by the fitted model. As shown in **Figure 4**, the growth of the Guinea deaths is exponential. The best fit curve for the projected growth is $y = 72.827e^{0.0823x}$. A comparison of the actual data to the projected data shows that the two are similar but not exact (**Table 2**). The projected amount of deaths is approximately 1300 by week 35 (or the week of November 23, 2014).

4.2. Modeling of Liberia Ebola deaths (weekly)

Unlike the Guinea deaths, the Liberian deaths are modeled using polynomial function (**Figure 5**).

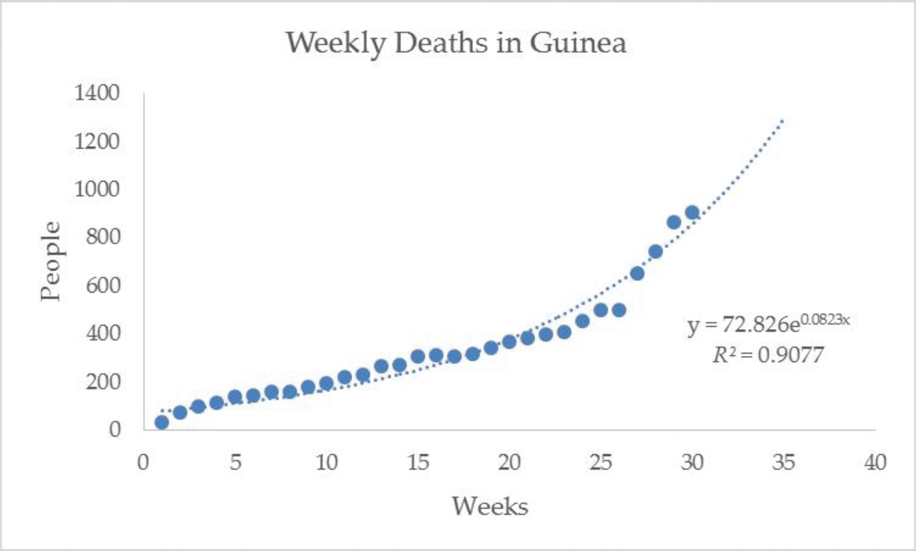

Figure 4. Weekly deaths in Guinea.

Ebola deaths in Guinea								
Week	**Deaths**	**Model**	**Week**	**Deaths**	**Model**	**Week**	**Deaths**	**Model**
1	29	79	13	264	212	25	494	570
2	70	86	14	267	231	26	494	619
3	95	93	15	303	250	27	648	672
4	108	101	16	307	272	28	739	730
5	136	110	17	304	295	29	862	792
6	143	119	18	314	320	30	904	860
7	155	130	19	339	348	31	XXX	934
8	157	141	20	363	378	32	XXX	1014
9	174	153	21	377	410	33	XXX	1101
10	193	166	22	396	445	34	XXX	1195
11	215	180	23	406	483	35	XXX	1298
12	226	196	24	450	525	36	XXX	XXX

Table 2. Actual and projected Ebola deaths in Guinea.

The best fit curve is best defined with the polynomial equation $y = 0.0003x^5 - 0.0069x^4 + 0.0347x^3 + 0.5074x^2 - 4.1442x + 10.487$. The model is not exact but it is close enough to predict that by week 35, there would be over 7000 deaths in Liberia (**Table 3**).

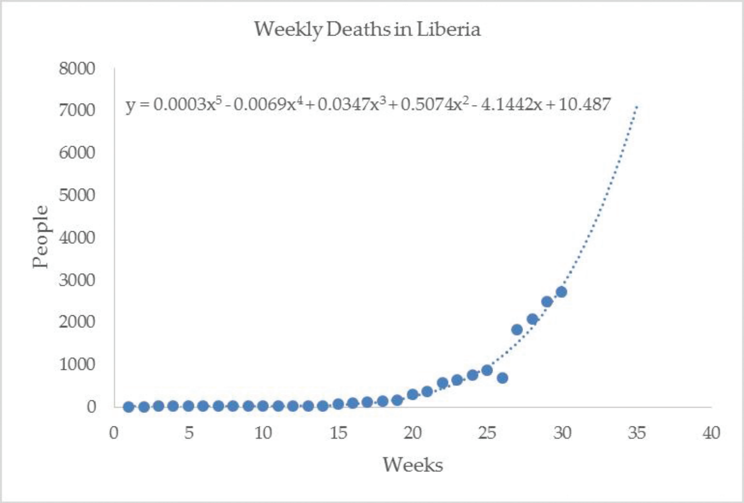

Figure 5. Weekly deaths in Liberia.

Ebola deaths in Liberia								
Week	**Deaths**	**Model**	**Week**	**Deaths**	**Model**	**Week**	**Deaths**	**Model**
1	0	7	13	24	33	25	871	1001
2	0	4	14	25	43	26	670	1267
3	10	3	15	65	58	27	1830	1589
4	13	3	16	84	79	28	2069	1976
5	6	3	17	105	107	29	2484	2436
6	6	5	18	127	145	30	2705	2981
7	11	7	19	156	197	31	XXX	3620
8	11	9	20	282	264	32	XXX	4366
9	11	12	21	355	352	33	XXX	5231
10	11	15	22	576	464	34	XXX	6230
11	11	20	23	624	606	35	XXX	7377
12	11	25	24	748	783	36	XXX	XXX

Table 3. Actual and projected Ebola deaths in Liberia.

4.3. Modeling of total deaths (World)

When analyzing the total deaths of Ebola (for 35 weeks), the data was best modeled using the polynomial function $y = 0.033x^4 - 1.4617x^3 + 23.437x^2 - 118.18x + 231.59$ (**Figure 6**). An exponential function was not used as it was not suitable since the actual growth was not (initially) fast enough to match the exponential growth. As shown in **Table 4**, the projected total deaths according to this model would be greater than 11,000 by week 35.

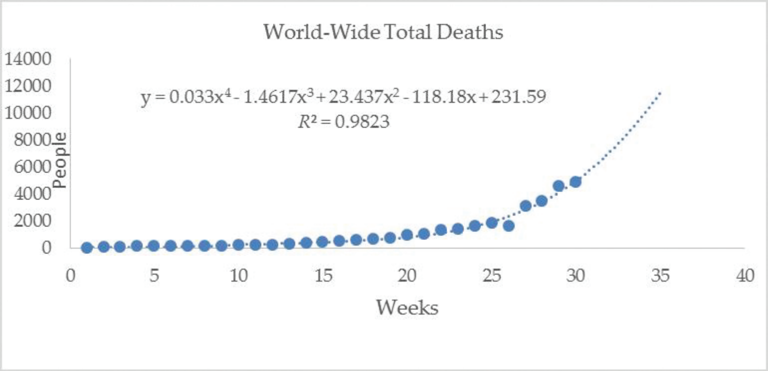

Figure 6. Weekly world-wide deaths.

Total Ebola deaths in the world

Week	Deaths	Model	Week	Deaths	Model	Week	Deaths	Model
1	29	135	13	337	387	25	1848	1977
2	70	78	14	350	428	26	1647	2392
3	105	51	15	467	470	27	3091	2893
4	121	49	16	518	516	28	3439	3494
5	142	65	17	603	571	29	4555	4206
6	149	93	18	660	638	30	4877	5044
7	166	131	19	729	722	31	XXX	6022
8	168	173	20	932	829	32	XXX	7155
9	185	217	21	1069	967	33	XXX	8461
10	210	262	22	1350	1141	34	XXX	9955
11	232	305	23	1427	1362	35	XXX	11656
12	244	347	24	1638	1637	36	XXX	XXX

Table 4. Actual and projected worldwide deaths.

4.4. Nonlinear regression and *R*-squared analysis

A visual inspection of the graphs and tables shows that the model for Liberia as well as the model for the world-wide total deaths evidently fits the data more closely and a lot better than does the Guinea model. Hence, other statistical goodness-of-fit tests are used to reassert these observations. Here, nonlinear polynomial regression (Eq. (11)) and R^2 statistical analysis are employed. In Eq. (11), Σ signifies summation, *w* refers to the actual (observed) number of Ebola deaths, *z* is the number of Ebola deaths as calculated with the model and *n* is the total number of weeks.

$$R^2 = \frac{2\left(\sum wz\right) + n(\bar{w})^2 - \sum\left(z^2\right) - 2\bar{w}\left(\sum w\right)}{\sum\left(w^2\right) + n(\bar{w})^2 - 2\bar{w}\left(\sum w\right)} \tag{11}$$

For the Guinea epidemiological Ebola model, the nonlinear regression equation is $y = 72.827e^{0.0823x}$ with R^2 as 0.9077 indicating that about 91% of the total variations in *y* (the number of actual Ebola deaths) can be explained by the regression equation. The polynomial epidemiological model for Ebola deaths in Liberia, $y = 0.0003x^5 - 0.0069x^4 + 0.0347x^3 + 0.5074x^2 - 4.1442x + 10.487$, has R^2 as 0.9715 so that about 97% of the total variations in *y* (the number of observed Ebola deaths) can be explained by the regression equation. For the third worldwide model, the polynomial for the total Ebola deaths for all countries combined is expectedly better. Here, the R^2 is 0.9823, so that about 98% of the total variations in the number of actual Ebola deaths can be explained by the regression equation, $y = 0.033x^4 - 1.4617x^3 + 23.437x^2 - 118.18x + 231.59$.

This shows that recording good and organized data that is easily retrievable is paramount in the fight of pandemics. The statistical models developed, in turn, can continue to inform and drive current and subsequent pandemic analyses.

5. Probability and expected value in statistics

At Wesley College, probability and expected value in statistics are introduced in two freshman-level mathematics classes: the quantitative reasoning math-core course and a first-year seminar, *Mathematics in Gambling*.

In general, there are two practical approaches to assigning a probability value to an event:

a. The classical approach

b. The relative frequency/empirical approach and

The **classical approach** to assigning a probability assumes that all outcomes to a probability experiment are equally likely. In the case of a roulette wheel at a casino, the little rolling ball is equally likely to land in any of the 38 compartments of the roulette wheel. In general, the rule for the probability of an event according to the classical approach is:

$$P\ (event\ A)\ = \frac{number\ of\ ways\ event\ A\ can\ occur}{total\ number\ of\ ways\ anything\ can\ occur} \tag{12}$$

In the case of roulette, the probability an individual wins by placing a bet on the color red is 18/38. Since there are 18 red, 18 black and 2 green compartments, the probability of a gambler winning by placing a bet on the color red is $\frac{18}{38} = \frac{9}{19}$ or approximately 0.474.

Unfortunately, the classical approach to probability is not always applicable. In the insurance industry, actuaries are interested in the likelihood of a policyholder dying. Since the two events of a policyholder living or dying are not equally likely, the classical approach cannot be used.

Instead, the **relative frequency approach** is used, which is:

$$P\ (event\ B)\ = \frac{number\ of\ times\ event\ B\ has\ happened\ in\ the\ past\ n\ trials}{number\ of\ trials,\ \ n} \tag{13}$$

When setting life insurance rates for policyholders, life insurance companies must consider variables such as age, sex and smoking status (among others). Suppose recent mortality data for 65-year-old non-smoking males indicates 1800 such men died last year out of 900,000 such men. Based on this data, one would say the probability a 65-year-old non-smoking male will die in the next year, based on the relative frequency approach is:

P (65-year-old non-smoking male dies) = $\frac{1,800}{900,000}$ or approximately 0.002 or 0.2%.

The field of decision analysis often employs the concept of **expected value**. Take the case of a 65-year-old non-smoking male buying a $250,000 term life insurance policy. Is it worth buying this policy? Based on the concept of expected value, a calculation based on probability is made and interpreted. If the value turns out to be negative, students then have to explain the rationale justifying the purpose of purchasing the term life insurance policy.

For a casino installing, a roulette wheel or craps table will the table game be a money maker for the casino? In the *Mathematics of Gambling* first-year seminar course, students research the rules for the game of roulette and the payoffs for various bets. Based on their findings, they determine the "house edge" for various bets. They also compare various bets in different games of chance to analyze which is a "better bet" and in what game.

Assume a situation has various outcomes/states of nature which occur randomly and are unknown when a decision is to be made. In the case of a person considering a life-insurance policy, the person will either live (L) or die (D) during the next year. Assuming the person has no adverse medical condition, the person's state of nature is unknown when he has to make the decision to buy the term life-insurance (the two outcomes will occur in no predictable manner and are considered random). If each monetary outcome (denoted O_i) has a probability (denoted p_i), then the **expected value** can be computed by the formula:

$$\begin{aligned} Expected\ \ Value\ &= O_1 \cdot p_1 + O_2 \cdot p_2 + O_3 \cdot p_3 + O_4 \cdot p_4 + \ldots + O_n \cdot p_n \\ &= \sum_{i=1}^{n} (O_i \cdot p_i) \end{aligned} \tag{14}$$

where there are n possible outcomes.

In other words, it is the sum of each monetary outcome times its corresponding probability.

Example 1: A freshman-level quantitative reasoning mathematics-core class

Assume a 67-year-old non-smoking male is charged \$1180 for a one year \$250,000 term life-insurance policy. Assume actuarial tables show the probability of death for such a person to be 0.003. What is the expected value of this life-insurance policy to the buyer?

A payoff table can be constructed showing the outcomes, probabilities and "net" payoffs:

Outcome:	Person dies	Person lives
Probability:	0.003	1 – 0.003 = 0.997
Net payoff:	\$250,000–\$1180	– \$1180
	\$248,820	

The payoff in the case of the person living is negative since the money is spent with no return on the investment. Using these data, the expected value is calculated as

$$Expected\ \ Value\ = \ \$248{,}820 \cdot 0.003 + -\$1{,}180 \cdot 0.997 = -\$430. \tag{15}$$

The negative sign in the expected value means the consumer should expect to lose money (while the insurance company can expect to make money). Students are asked to explain the meaning of the expected value and explain reasons for people throwing their money away like this. What will they do when it comes time to consider term life insurance?

Example 2: Mathematics of Gambling class

Students are asked to research rules of various games of chance, the meaning of various payoffs (for example, 35 to 1 versus 35 for 1) and then be asked to calculate and interpret the **house edge** in gambling. This is defined by the formula

$$House\ Edge\ = \ \frac{Expected\,Value\,of\ the\,Bet}{Size\,of\ the\,Bet} \tag{16}$$

By asking different students to evaluate the house edge of different gambling bets, students can analyze and decide which bet is safest if they do choose to gamble.

Which bet has the lower house edge and why?

Bet #1 – Placing a $10 bet in American roulette on the "row" 25– 27.

Bet #2 – Placing a $5 bet in Craps on rolling the sum of 11.

Students must research each game of chance and determine important information to use, which is recorded as follows:

	$10 Bet on a row in roulette	$5 Bet on a sum of 11 in craps
Probability of a winning bet:	$\frac{3}{38}$	$\frac{2}{36} = \frac{1}{18}$
Payoff odds:	11 to 1	15 to 1
Payoff:	−$110	−$75
Probability of a losing bet:	$\frac{35}{38}$	$\frac{34}{36} = \frac{17}{18}$
Payoff to house for lost bet:	+$10	+$5
House Edge:	$0.0526	$0.1111
Computed by:	$\dfrac{\frac{3}{38} \cdot (-\$110) + \frac{35}{38} \cdot (+\$10)}{\$10}$	$\dfrac{\frac{1}{18} \cdot (-\$75) + \frac{17}{18} \cdot (+\$5)}{\$5}$

The roulette bet has a lower house edge and is financially safer in the long run for the gambler. Students were then asked to compute the house edge using the shortcut method based on the theory of odds. The house edge is the difference between the true odds (denoted *a:b*) and the payoff odds the casino pays, expressed as a percentage of the true total odds ($a + b$).

In the example involving craps, the true odds against a sum of 11 is 34:2 which reduces to 17:1. The difference between the true odds and payoff odds is 17 – 15 (see Example 2) = 2. Expressing this difference as a percentage of (a + b), the house edge is then calculated as $2 \div (17 + 1) = 2 \div 18 = \frac{1}{9} = 0.1111$ which is the same answer found using the expected value.

Due to the concept of the house edge, casinos know that in the long run, every time a bet is made in roulette, the house averages a profit of $0.0526 for each dollar bet. Yes, gamblers do win at the roulette table and large amounts of money are paid out. But in the long run, the game is a money maker for the casino.

Acknowledgements

This work was made possible by grants from the National Institute of General Medical Sciences —NIGMS (P20GM103446) from the National Institutes of Health (DE-INBRE IDeA program), a National Science Foundation (NSF) EPSCoR grant IIA-1301765 (DE-EPSCoR program) and the Delaware (DE) Economic Development Office (DEDO program). The undergraduates acknowledge tuition scholarship support from Wesley's NSF S-STEM Cannon Scholar Program (NSF DUE 1355554) and RB acknowledges further support from the NASA DE-Space Grant Consortium (DESGC) program (NASA NNX15AI19H). The DE-INBRE, the DE-EPSCoR and the DESGC grants were obtained through the leadership of the University of Delaware and the authors sincerely appreciate their efforts.

Author contributions

Drs. D'Souza, Wentzien and Nwogbaga served as undergraduate research mentors to Brandenberg, Bautista and Miller, respectively. Professor Olsen has developed and taught the probability and expected value examples in his freshman-level mathematics core courses. The findings and conclusions drawn within the chapter in no way reflect the interpretations and/or views of any other federal or state agency.

Conflicts of interest

The authors declare no conflict of interest.

Author details

Malcolm J. D'Souza[1*], Edward A. Brandenburg[1], Derald E. Wentzien[2], Riza C. Bautista[2], Agashi P. Nwogbaga[2], Rebecca G. Miller[2] and Paul E. Olsen[2]

*Address all correspondence to: malcolm.dsouza@wesley.edu

1 Department of Chemistry, Wesley College, Dover, Delaware, USA

2 Department of Mathematics and Data Science, Wesley College, Dover, Delaware, USA

References

[1] D'Souza, M.J., Curran, K.L., Olsen, P.E., Nwogbaga, A.P., Stotts, S. Integrative Approach for a Transformative Freshman-Level STEM Curriculum. Journal College Teaching & Learning, 2016; 13:47–64.

[2] D'Souza, M.J., Kashmar, R.J., Hurst, K., Fiedler, F., Gross, C.E., Deol, J.K., Wilson, A. Integrative Wesley College Biological Chemistry Program Includes the Use of Informatics Tools, GIS and SAS Software Applications. Contemporary Issues in Education Research, 2015; 8:193–214.

[3] Kevill, D.N., D'Souza, M.J. Sixty Years of the Grunwald-Winstein Equation: Development and Recent Applications. Journal of Chemical Research, 2008; 61–66.

[4] D'Souza, M. J., Kevill, D.N. Application of the Grunwald-Winstein Equations to Studies of Solvolytic Reactions of Chloroformate and Fluoroformate Esters. Review Chapter in Research Recent Developments in Organic Chemistry, 2013; 13:1–38, Editor, S.G. Pandalai. Transworld Research Network, Kerala, India, ISBN: 978-81-7895-600-8.

[5] Kevill, D.N., D'Souza, M.J. Correlation of the Rates of Solvolysis of Benzoyl Fluoride and a Consideration of Leaving-Group Effects. Journal of Organic Chemistry, 2004; 69:7044–7050.

[6] Kevill, D.N., D'Souza, M.J. Correlation of the Rates of Solvolysis of Benzoyl Chloride and Derivatives Using Extended Forms of the Grunwald-Winstein Equation. Journal of Physical Organic Chemistry, 2002; 15:881–888.

[7] Kim, J.W., Lee, I., Sohn, S.C., Uhm, T.S. Solvolysis of Benzoyl Cyanide, Journal of the Korean Chemical Society, 1983; 27:95–101.

[8] Centers for Disease Control and Prevention (2016): 2014 Ebola Outbreak in West Africa. Retrieved from http://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/index.html.

[9] World Health Organization. Global Alert and Response: Ebola Virus Disease (EVD). Geneva, Switzerland: World Health Organization; 2014. Available at http://www.who.int/csr/don/archive/disease/ebola/en.