www.simiode.org
SIMIODE
A SYSTEMIC INITIATIVE FOR MODELING INVESTIGATIONS & OPPORTUNITIES WITH DIFFERENTIAL EQUATIONS

STUDENT VERSION
Gender Demographics in Engineering

Brody Johnson, Elodie Pozzi
Department of Mathematics & Statistics
Saint Louis University
Saint Louis MO USA

SCENARIO DESCRIPTION

In 1966, fewer than 1% of the 35,826 bachelor's degrees in engineering awarded in the United States were earned by women [1]. This percentage showed rapid growth over the following ten years, but undergraduate women in the United States still remain underrepresented in engineering some fifty years later, despite comprising the majority of the undergraduate student body. In fact, women comprised 57% of the undergraduate student body in 2017 while earning just 22% of the bachelor's degrees in engineering awarded that year [2, 1]. This project makes no attempt to explain why women remain underrepresented in undergraduate engineering programs; however, students are encouraged to learn more about this issue by visiting the Society of Women Engineers website (https://swe.org/) or reading about the history of this organization [3]. Rather, the goal of this project is to use historical data to examine the following question:

> *Does the growth in the percentage of women in undergraduate engineering programs over the last fifty years appear to be on track to achieve a gender balance mirroring that of the entire undergraduate population?*

Towards this end, a mathematical model will be developed for the percentage of bachelor's degrees in engineering earned by women in the United States as a function of time, based on real data. The data comes from the American Physical Society (https://www.aps.org/) and includes the percentage of bachelor's degrees in engineering earned by women in the United States for each year between 1966 and 2018 [1]. The time history is shown in Figure 1.

Let $P(t)$ represent the percentage of bachelor's degrees in engineering that are awarded to women in year t. It will be assumed that $P(t)$ obeys a differential equation of the form

$$\frac{dP}{dt} = f(P), \tag{1}$$

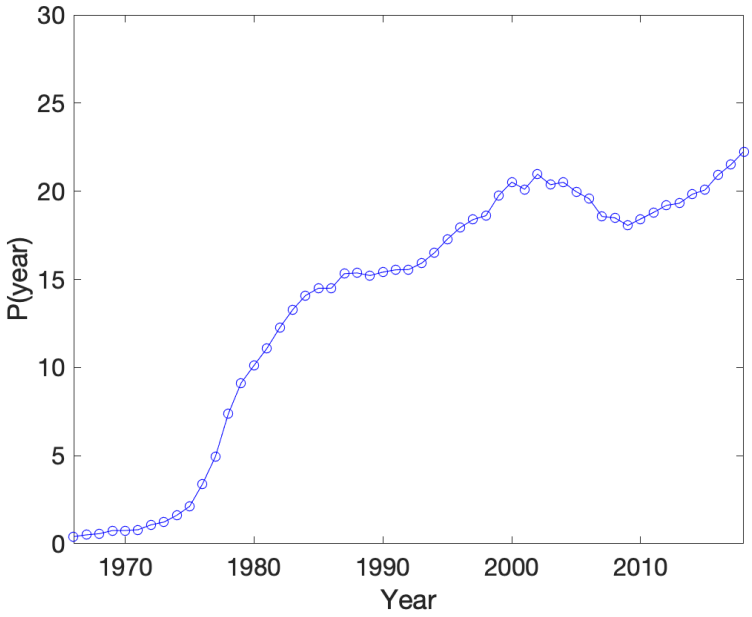


Figure 1. Percentage of Bachelor's degrees in engineering earned by women in the United States [1].

where f is a continuous function. As a percentage, the function $P(t)$ is confined to the range $0 \leq P(t) \leq 100$ and thus it is reasonable to employ a population model that accounts for limited resources. The *growth function* $f(P)$ for such models usually satisfies the following general properties:

- $f(0)$ should be zero;
- $f(P)$ should be positive for small positive values of P;
- $f(P)$ should be negative for sufficiently large values of P.

These assumptions reflect the fact that a small population grows steadily due to an abundance of resources, but growth will slow and possibly recede as the resources are exhausted. Although the above criteria could be met by countless varieties of growth functions, two choices tend to receive the most attention. The first of these is the so called *logistic* growth function,

$$f(P) = kP\left(1 - \frac{P}{N}\right), \qquad k,\, N > 0. \tag{2}$$

The second is called the *Gompertz* growth function and is given by

$$f(P) = kP\ln\left(\frac{N}{P}\right), k,\, N > 0. \tag{3}$$

One further growth function will be introduced for consideration. This growth function will be referred to as the *cubic* growth function and is described by

$$f(P) = kP\left(1 - \frac{P}{N}\right)^3, k,\ N > 0. \tag{4}$$

Figure 2 offers a graphical comparison of the three different types of growth functions, each with $N = 50$. (The choice $N = 50$ is used to reflect a population that is divided evenly between men and women.) Do these functions satisfy the three general properties described above? How will the differences in the shapes of curves influence the corresponding models?

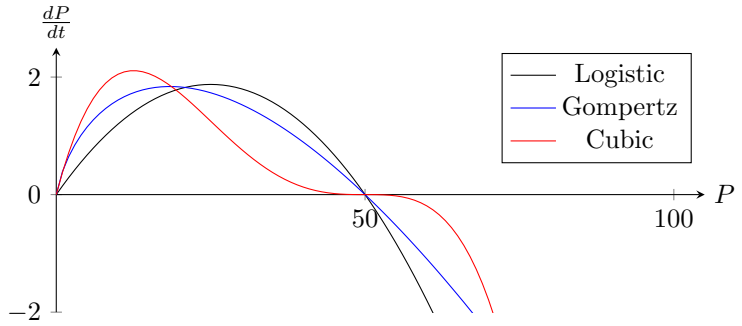


Figure 2. Comparison of the Logistic, Gompertz, and Cubic Population Models.

Ultimately, these models will be used in an attempt to forecast the growth of the actual percentage of undergraduate engineering degrees awarded to women in the United States over the next five to ten years. To accomplish this, it will be important to understand how the constants k and N affect the shape of the graph for $P(t)$. It will also be important to recognize differences between the three models and determine which is best suited to the data. Once this has been done, it will be possible to approach the main question by generating a forecast for the percentage of engineering degrees awarded to women in the future. The remainder of this modeling scenario will break this large task into a series of steps, many of which rely on knowledge developed earlier in the course.

1 Qualitative Analysis

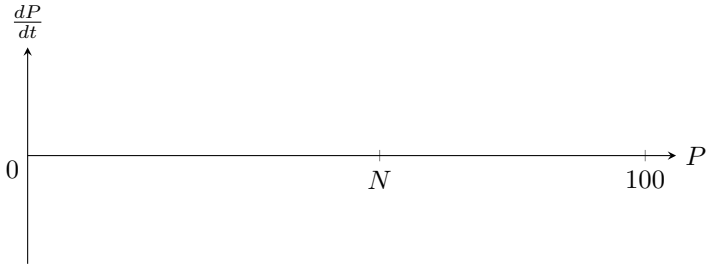
Qualitative analysis leads to the understanding of certain characteristics of the solutions of a differential equation (what is going to happen) without explicitly solving the differential equation (no solution formula is used). This is especially important when the differential equation cannot be solved using available techniques. Notice that the differential equation (1) is *autonomous* for any growth function $f(P)$, which means the slope $\frac{dP}{dt}$ at any point (t, P) is dependent only on P. Recall that the *phase line* of an autonomous differential equation $\frac{dP}{dt} = f(P)$ is essentially an annotated graph of $f(P)$ versus P. Positive values of $f(P)$ correspond to positive values of $\frac{dP}{dt}$, which means P will increase (move to the right) as t increases. Similarly, negative values of $f(P)$ correspond to

negative values of $\frac{dP}{dt}$, which means P will decrease (move to the left) as t increases. These directions are noted on the P-axis using arrows to the right or left. If $f(P) = 0$, however, then $\frac{dP}{dt} = 0$ and $P(t)$ will remain constant. Such a point P is called an *equilibrium point* of the differential equation. If nearby solutions tend towards the equilibrium point it is said to be *stable*. Otherwise, the equilibrium point is said to be *unstable*.

Student Task: Construct a phase line for the cubic model,

$$\frac{dP}{dt} = kP\left(1 - \frac{P}{N}\right)^3, \qquad (0 < N < 100)$$

in the space provided below. Include arrows on the P-axis that illustrate the direction of solutions. Be sure to identify any equilibrium points and label each as stable or unstable.



Questions:

1.1 How many equilibrium points does this model have? Describe the stability of each equilibrium point.

1.2 If $0 < P(0) < 100$, what can you say about $P(t)$ for large values of t? Try to interpret your answer in the context of the percentage of women earning undergraduate degrees in engineering.

1.3 How do the parameters k and N affect the phase line? Does one parameter have a more noticeable effect?

1.4 Would the phase line or any of your subsequent answers change if you replace the cubic model with the logistic model or the Gompertz model? Explain.

1.5 How do the logistic, Gompertz, and cubic growth functions differ? How do you think these differences will affect the solution curves for these models? (You may want to graph each of the growth functions on a common plot.)

2 Numerical Solution

Regardless of the choice of the growth function, notice that (1) is always separable, so one can attempt to find an explicit formula for the solutions via integration. This works well for the logistic and Gompertz models, which is one of the reasons these models are often used. The cubic model,

on the other hand, is not easily solved in this manner. In this case it makes sense to implement a numerical method to obtain an approximate solution of the differential equation. The improved Euler method, summarized in Figure 3, will be adopted for this project.

Improved Euler Method:
Choose $\Delta t > 0$ and consider the initial value problem

$$\frac{dP}{dt} = f(P) \quad \text{subject to} \quad P(t_0) = P_0.$$

Define $t_j = t_0 + j\Delta t$. The approximate value of $P(t_j)$, denoted by P_j, can be computed from P_{j-1} as follows:

1. Compute the initial slope estimate: $m_1 = f(P_{j-1})$.
2. Compute a temporary estimate for P: $\tilde{P} = P_{j-1} + m_1\Delta t$.
3. Compute the second slope estimate: $m_2 = f(\tilde{P})$.
4. Compute P_j using the average slope:

$$P_j = P_{j-1} + \left(\frac{m_1 + m_2}{2}\right)\Delta t.$$

Figure 3. The improved-Euler method.

Student Task: Create a spreadsheet to implement the improved Euler method for the cubic model to solve the initial value problem

$$\frac{dP}{dt} = kP\left(1 - \frac{P}{N}\right)^3 \quad \text{subject to} \quad P(0) = 1$$

over the range $0 \le t \le 52$. Use $\Delta t = 1$, $k = 0.2$, and $N = 57$.

Tips for the Spreadsheet: (see Figure 4)

- Create six columns: (A) Iteration j, (B) Time t_j, (C) Population P_j, (D) Slope $m_1 = f(P_j)$, (E) Temporary Estimate $\tilde{P}$, (F) Slope $m_2 = f(\tilde{P})$.
- Create cells for the parameters k and N and reference these cells in the two slope formulas.
- Add a scatter plot to graph P_j as a function of t_j.

	A	B	C	D	E	F
1	Cubic Model:		dP/dt = k P(1-P/N)^3			
2	Parameters	k	N1	P0	SSE	delta t
3		0.1000	57.00	1.0000	1563.45	1
4						
5		(years)	Pj	m1	y-tilde	m2
6	Iteration	time	Model	slope	temp	slope-2
7	0	0	1.000000	0.094829	1.094829	0.103295
8	1	1	1.099062	0.103670	1.202732	0.112819

Figure 4. Screenshot of a sample spreadsheet.

Questions:

2.1 How long does it take before P_j exceeds 25? How sensitive is this time to the changes in te parameters k and N?

2.2 How does the graph of the approximate solution change when k is increased? What if k is decreased?

2.3 Create new sheets to implement the logistic and Gompertz models by copying and modifying the current sheet. What changes are required? How do the solution curves differ?

2.4 Does the data in Figure 1 exhibit any characteristics of the solution curves found here? Explain.

3 Model Implementation

The last step in this project involves the application of the three population models to actual data for the percentage of bachelor's degrees in engineering earned by women in the United States between 1966 and 2018. The parameters k and N will be adjusted so that each model predicts the data as accurately as possible. The previous steps should have provided some insight about how changes to the parameters k and N will influence the shape of the solution predicted by the model. This insight will help determine whether a given value of k or N in the model is too small or too large based on the comparison with the historical data. For convenience, the raw data depicted in Figure 1 is presented as Table 1. This data is also available in a spreadsheet among the supporting documents.

Student Task: Modify your spreadsheet to compare the observed data to predicted values of $P(t)$ for any values of the parameters k and N. Optimize the values of k and N for each model so that the sum of the squared error (SSE) is minimized. The sum of the squared error is computed as follows:

$$\text{SSE} = \sum_{j=1}^{J} \left(P_j - P(t_j)\right)^2,$$

where J is the number of data points.

Year	P(Year)	Year	P(Year)	Year	P(Year)	Year	P(Year)
1966	0.4075	1980	10.1207	1994	16.5096	2008	18.4923
1967	0.5083	1981	11.0850	1995	17.2792	2009	18.0604
1968	0.5732	1982	12.2665	1996	17.9368	2010	18.4281
1969	0.7503	1983	13.2820	1997	18.3960	2011	18.7971
1970	0.7527	1984	14.0887	1998	18.6147	2012	19.1987
1971	0.7978	1985	14.4975	1999	19.7700	2013	19.3157
1972	1.0763	1986	14.4988	2000	20.5188	2014	19.8257
1973	1.2313	1987	15.3228	2001	20.1118	2015	20.0755
1974	1.6139	1988	15.3648	2002	20.9667	2016	20.9124
1975	2.1218	1989	15.2180	2003	20.3679	2017	21.5032
1976	3.3952	1990	15.4130	2004	20.5111	2018	22.2550
1977	4.9423	1991	15.5418	2005	19.9740		
1978	7.3692	1992	15.5567	2006	19.5750		
1979	9.1287	1993	15.9174	2007	18.5692		

Table 1. Percentage of Bachelor's degrees in engineering earned by women in the United States [1].

Tips for the Spreadsheet:

- Add two additional columns: (G) Actual Data $P(t_j)$, (H) Squared Error.
- Interpret 1966 as $t = 0$ so that 2018 corresponds to $t = 52$.
- Modify P_0 to match the starting value of the data.
- Create a cell for the sum of the squared error column.
- Add a line or scatter plot to compare $P(t_j)$-observed with P_j-predicted.
- Start with $k = 0.1$ and $N = 50$. Make small adjustments to k or N while watching the graphs and the sum of the squared error. Think about how the changes will affect the graph of $P(t)$-predicted before implementing them.
- If the spreadsheet is equipped with Solver, use the Solver to optimize the values of k and N so that the sum of the squared errors is minimized.
- Repeat these steps for each of the three models.

Questions:

3.1 Do any of the models fail to capture important trend(s) in the data? Explain.

3.2 What do the optimal values of k and N convey about the progress towards a gender balance in the engineering sciences through 2018? Do the models agree in this regard?

3.3 Which model led to the smallest value for the sum of the squared error?

3.4 Which model do you think will predict $P(t)$ the most accurately over the next ten years?

REFERENCES

[1] American Physical Society. https://www.aps.org/programs/education/statistics/womenmajors.cfm. Accessed 9 June 2021.

[2] National Council for Educational Statistics. https://nces.ed.gov/fastfacts/display.asp?id=98. Accessed 9 June 2021.

[3] Wikipedia article: 'Society of Women Engineers'. https://en.wikipedia.org/wiki/Society_of_Women_Engineers. Accessed 16 June 2021.