classic

The Midge Problem

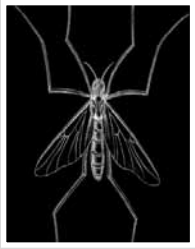
GARY FROELICH

DANIEL TEAGUE AND HELEN COMPTON

This is a reprint of the first Everybody's Problems column with two new solutions from the authors. They discuss a problem from the 1989 MCM, in which a team from their high school, the North Carolina School of Science and Mathematics (NCSSM), submitted a Meritorious paper. Although COMAP now runs an undergraduate modeling competition (MCM) and a high school modeling competition (HiMCM), NCSSM still enters the undergraduate competition and regularly does very well. Dan and Helen will tell you that their students succeed on MCM because modeling is an everyday concern in the mathematics courses at NCSSM. Until the same can be said of all American high schools, columns like this cannot be printed too often. ❑

In 1989, the Mathematical Contest in Modeling offered a wonderful problem about distinguishing a "good" midge from a "bad" midge. It is a fine example of a problem that can be used to excellent effect with students at many different levels. We have been giving a version of this problem to our Precalculus students for the past few years. Typically, the problem is presented after we have spent time studying techniques of data analysis and linear curve fitting. A unique aspect to this problem, as you'll see, is rather than fit a line to a set of data, the students are asked to use what they've learned to fit a line to the *absence* of data! What follows is the statement of the problem and five different approaches that student groups have used in solving it.

The Midge Problem



In 1981, two new varieties of a tiny biting insect called a midge were discovered by biologists W. L. Grogan and W. W. Wirth in the jungles of Brazil. They dubbed one kind of midge an **Apf** midge and the other an **Af** midge. The biologists found out that the **Apf** midge is a carrier of a debilitating disease that causes swelling of the brain when a human is bitten by an infected midge. Although the disease is rarely fatal, the disability caused by the swelling may be permanent. This is no insect to mess with! The other form of the midge, the **Af**, is quite harmless and a valuable pollinator. In an effort to distinguish the two varieties, the biologist took measurements on the midges they caught. The two measurements taken (in centimeters) were of wing length and antenna length.

Af Midges

Wing Length (cm)	1.72	1.64	1.74	1.70	1.82	1.82	1.90	1.82	2.08
Antenna Length (cm)	1.24	1.38	1.36	1.40	1.38	1.48	1.38	1.54	1.56

Apf Midges

Wing Length (cm)	1.78	1.86	1.96	2.00	2.00	1.96
Antenna Length (cm)	1.14	1.20	1.30	1.26	1.28	1.18

Is it possible to distinguish an **Af** midge from an **Apf** midge on the basis of wing and antenna length?
Write a report that describes to a naturalist in the field how to classify a midge he or she has just captured.

Sample Student Solutions

The students noticed that both data sets overlapped, so knowing only one of the measures did little to distinguish the midges. Of course, the first thing to do is graph the data, as shown in **Figure 1**.

From the graph, we can see that there is a clear region between the two data sets. Where should the boundary be placed to most accurately distinguish the two species? Five solutions are presented. As you will see, the solutions are from groups of students in different classes and with different levels of mathematical preparation and sophistication.

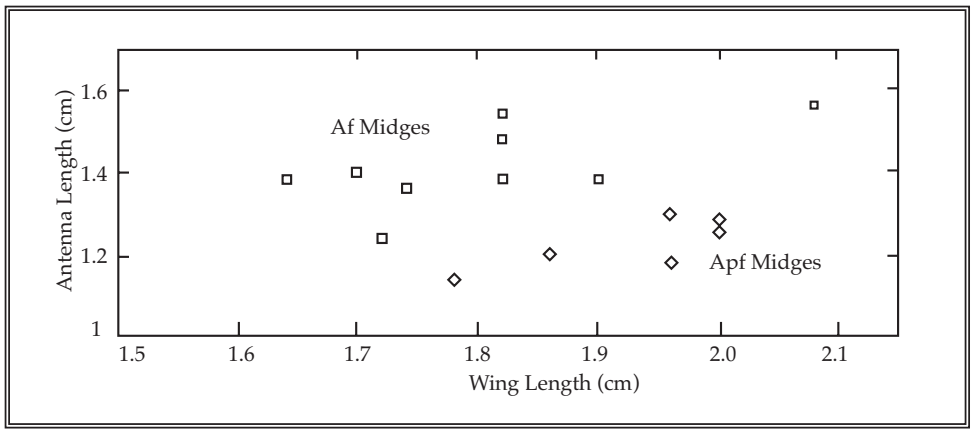


FIGURE 1. SCATTERPLOT OF THE MIDGE DATA.

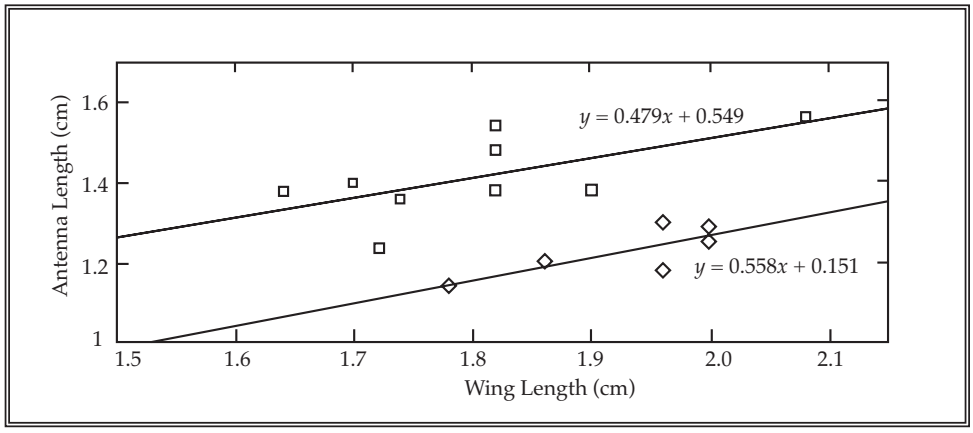


FIGURE 2. LINEAR LEAST-SQUARES LINE FIT TO EACH DATA SET.

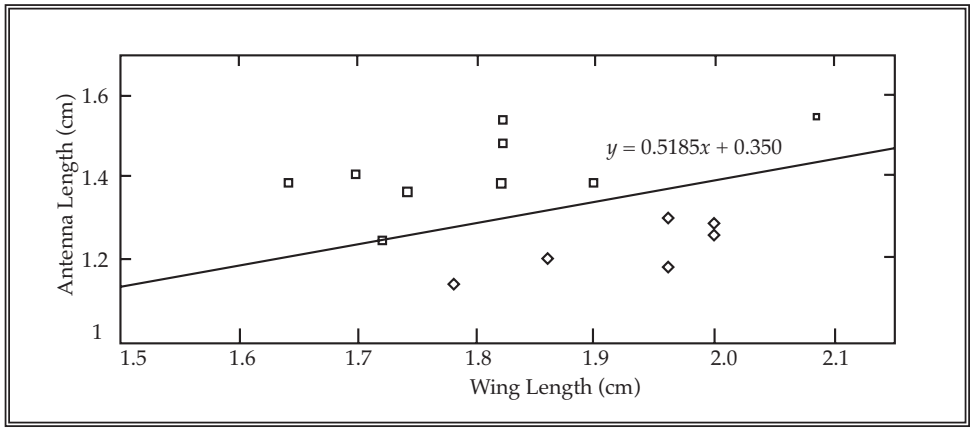


FIGURE 3. LEAST-SQUARES MID-LINE.

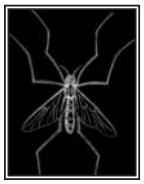
Solution 1:

The most common solution involved first fitting a least-squares or median-median line to each data set, using Wing Length as the independent variable and Antenna Length as the dependent variable. The least-squares linear fit for **Af** midges is $y = 0.479x + 0.549$ while for the **Apf** midges it is $y = 0.588x + 0.151$. (**Figure 2**.)

Since these two lines pass through the two data sets, it seemed reasonable that the mid-line between them would be a good boundary. To find the line that bisects the region between these two lines, simply "average" the two lines. The boundary determined in this fashion is $y = 0.5185x + 0.350$. Any midge below this line was to be considered an **Apf** midge and destroyed, while any midge above the line was to be considered an **Af** midge and saved. How does it look to you? (**Figure 3**.)

Most students felt that this line was not a good boundary, since it seems to misclassify an **Af** midge as an **Apf** midge. Almost all groups eventually realized another approach needed to be used. However, one group steadfastly argued that this misclassification was a reasonable price to pay for added safety. They would rather say a safe midge was dangerous, and erroneously kill it, than to say a dangerous midge was safe. This line, while clearly not a good choice if your interest is accuracy, is a good line to use if your interest is safety.

Solution 2:



Seeing the misclassification in the procedure above, some students used a different "averaging" process. They first found the two "outermost" midges in each group. These two midges define a line past which no midges of their species have been found. (**Figure 4**.)

For **Af** midges, this line is $y = 0.778x - 0.098$ while for **Apf** midges it is $y = 0.889x - 0.442$. As before, some students took the mid-line for the boundary, using $y = 0.8335x - 0.270$ while others wanted to err on the side of caution and used the **Af** boundary line $y = 0.778x - 0.098$. One group used this line with a little bit of room added, $y = 0.778x - 0.10$. (See **Figure 5**.)

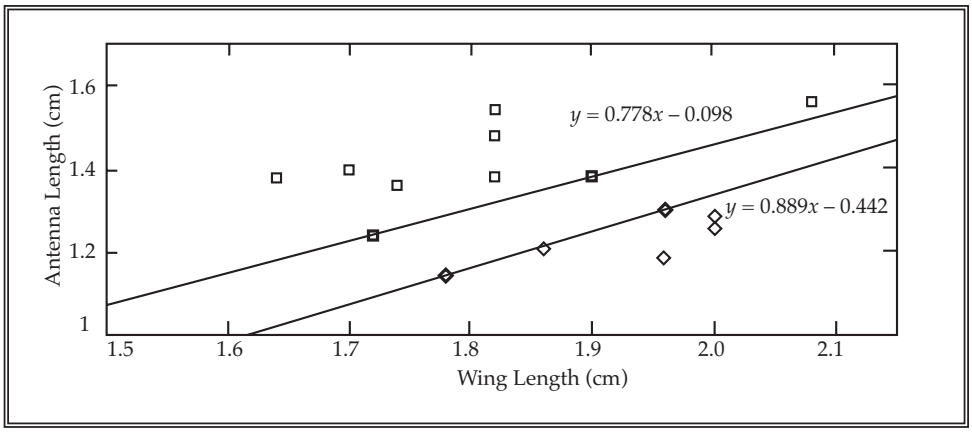


FIGURE 4. OUTERMOST MIDGE LINES.

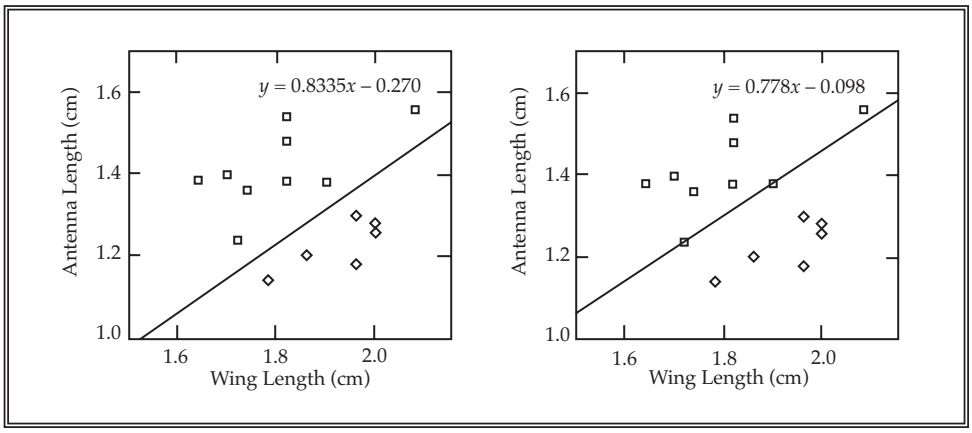
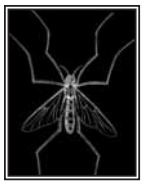


FIGURE 5. LINES FOR "ACCURACY" AND "SAFETY."

Solution 3:



One group of students argued that the **Af** midges seemed to have generally larger antennae and smaller wings. They thought that the ratio of antenna length to wing length might tell them something. The ratios are:

Af	0.721	0.841	0.782	0.824	0.758
	0.813	0.726	0.846	0.750	
Apf	0.640	0.645	0.663	0.630	0.640
	0.602				

Figure 6 shows these ratios on a number line.

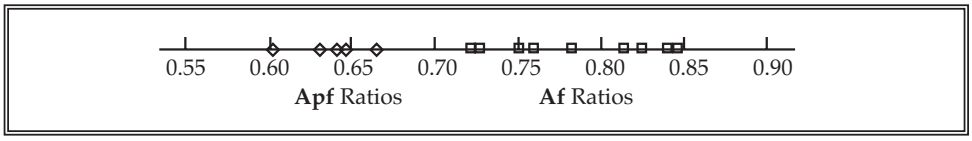


FIGURE 6. NUMBER LINE WITH ANTENNA-TO-WING-LENGTH RATIOS.

Notice that there is no overlap in these ratios. The smallest ratio for **Af** midges is 0.721 and the largest for **Apf** midges is 0.663. Groups differ on how to split up this interval [0.663, 0.721]. One group considered using the midpoint of this interval, so any midge with an antenna to wing ratio of less than 0.692 is considered an **Apf** midge. Other groups decided to "play it safe" and use the smallest **Af** value of 0.721 for their boundary. Others, recognizing that three-fifths of all the midges found have been **Af** midges, settled on a point three-fifths through the interval, by using $\left(\frac{3}{5}\right)(0.721) + \left(\frac{2}{3}\right)(0.663) = 0.698$. Any midge whose ratio is less than 0.698 is considered an **Apf** midge, and killed, while any midge whose ratio is larger than 0.698 is considered an **Af** midge, and saved. While their attempt was good, they had the proportion backwards. Since there are more **Af** midges, that part of the interval should be larger than that for the **Apf** midges. According to their argument, the students should have used 0.686 as their boundary. You might find other values in the interval [0.663, 0.721] to use.

Since the inception of the AP Statistics course, more and more students bring statistical knowledge into the mathematics classroom. A variation on the previous idea that utilizes statistical techniques is now a common approach. By using a normal probability plot (NPP), the

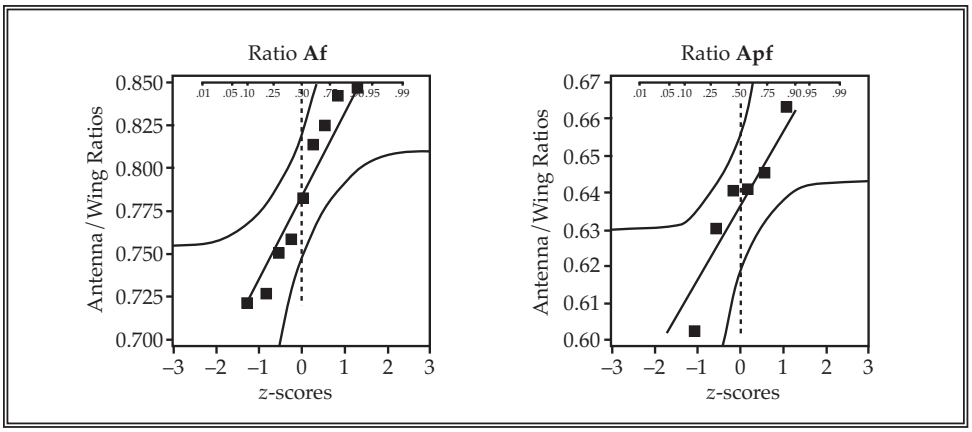


Figure 7. Normal Probability Plots.

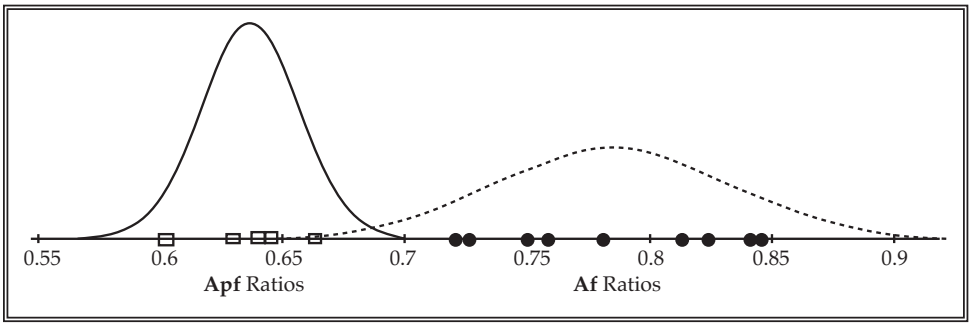


Figure 8. Approximate Normal Distributions for Ratios.

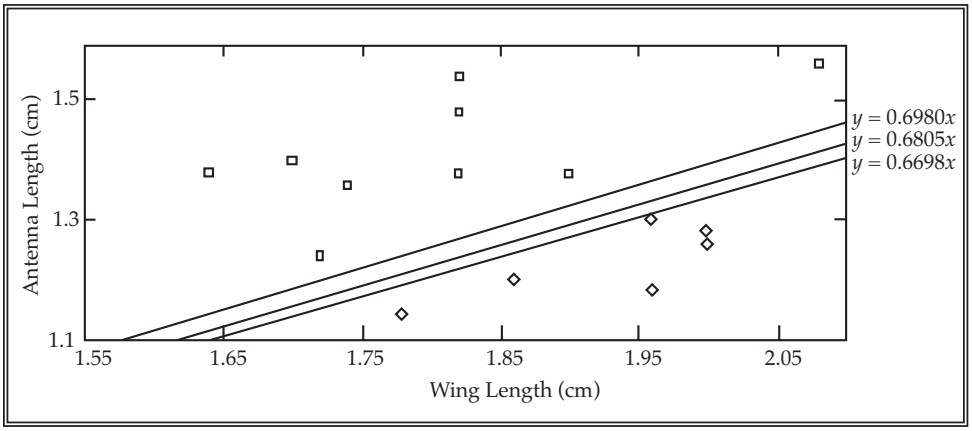


Figure 9. Linear boundary implied by Ratio Model.

distributions of the ratios can be shown to be approximately normal. The plots for the two midge ratios are shown in **Figure 7**.

Based on the NP plots, students can argue that the **Af** ratios are approximately normally distributed with a mean of 0.785 and standard deviation of 0.048, while the **Apf** ratios are approximately normally distributed with a mean of 0.637 and a standard deviation of 0.020. See **Figure 8**.

Since the **Apf** ratios ~$N(0.637, 0.020)$, students find that only 5% of **Apf** midges should have ratio larger than 0.637 + 1.645(0.020) = 0.6699. This is the boundary they choose. Midges with ratios less than 0.67 are considered **Apf** midges. We would expect to find less than 1% of the **Af** midges below this boundary. A safer choice would be to find the boundary where less than 1% of the **Apf** midges would fall above (this is 0.683), but the 5% criterion is very fixed in these students minds as the "best" value based on their experience in AP Statistics.

Other groups look for the boundary that would determine where the midge is equally likely to be **Af** or **Apf**. They do this by finding the location on the number line where the z-scores are the same for each group. A z-score, computed as $z = \frac{x-\bar{x}}{s}$, measures the distance from the mean in standard deviations. The students begin by solving the equation $\frac{x-0.637}{0.020} = \frac{x-0.785}{0.048}$. Unfortunately, the solution is 0.531, which cannot possibly work. After some consideration, some groups gave up and tried other approaches, but others realized that the z-score for the **Apf** midges will be positive, but the z-score for the **Af** midges will be negative. So the equation should be $\frac{x-0.637}{0.020} = \frac{x-0.785}{0.048}$. This solution is 0.6805, a solution that makes sense.

An interesting sidelight that the students failed to notice is that using the ratio is equivalent to using a line. If $\frac{y}{x} = 0.698$, then $y = 0.698x$. (See **Figure 9**.)

Solution 4:

Several other groups also noted the disparity in the number of midges in each group. Since three-fifths of the known midges are **Af** midges, they thought more "room" should be allowed for this group. They modified the results of Solutions 1 and 2 to accommodate this idea. Rather than split the region between the two lines in half, using the

mid-line, as in solutions 1 and 2, they wanted to put the line three-fifths of the way through. Without exception, the students have argued that since the **Af** midges needed three-fifths of the room, they multiplied the equation of the **Af** line by three-fifths and the equation of the **Apf** line by two-fifths. For Solution 1, this gives $y = 0.5106x - 0.3898$ and for Solution 2, $y = 0.8224x - 0.2356$. This, of course, is the same mistake made by the group using ratios. (**Figure 10**.)

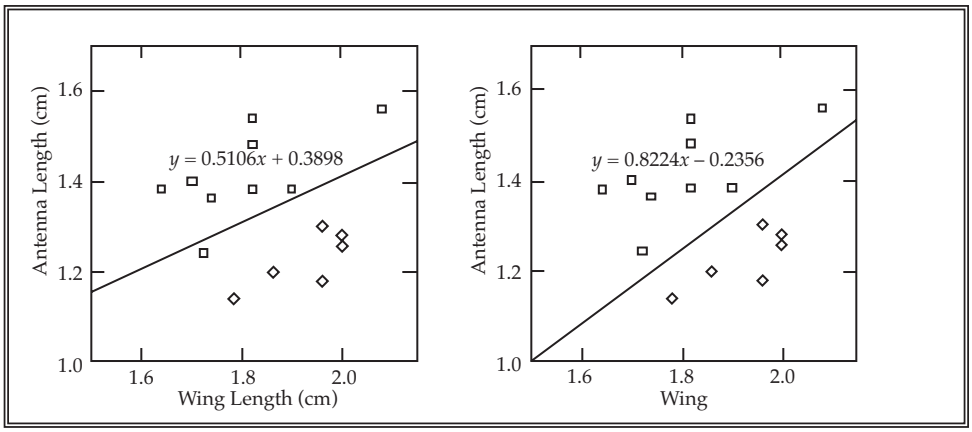


FIGURE 10. INCORRECT BOUNDARIES DUE TO WEIGHTING ERROR.

Those groups modifying Solution 1 realized that the weighting is backwards because of the misclassification. Unfortunately, those modifying Solution 2 did not realize that the process they used did not match the argument they gave. Well, that's a teaching point. The two lines should be $y = 0.5264x + 0.3102$ and $y = 0.8446x - 0.3044$. (See **Figure 11**.)

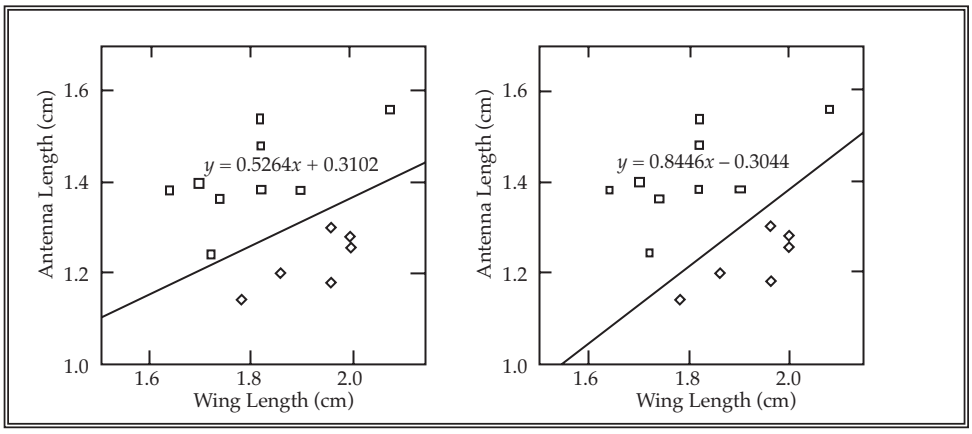
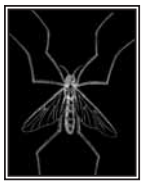


FIGURE 11. CORRECTLY WEIGHTED BOUNDARIES.

Solution 5:



The final solution, from a group in which two members were also taking Statistics, is the most sophisticated.

The question of whether there are more **Af** midges than **Apf** midges, since 9 of the 15 midges in the sample are **Af**, can be answered in several ways. The way to think about this statistically is to ask the question, "If there are the same number of each midge, and 15 were selected at random, how likely is it that you would select 9 or more of one type?" The students used their calculators to choose 0 or 1 at random, each with probability one-half. They then choose 15 times and found the sum. They repeated this experiment 50 times each, counting the number of times the sum was 9 or more. Of the 150 total trials (there were three in the group) the sum was 9 or more 41 times, or on 27% of the trials. Having 9 out of 15 midges be **Af** midges does not argue that there are more of this species.

A probabilistic argument is to compute the probability of getting 9 or more **Af** midges from a population that is evenly distributed between **Af** and **Apf**. This computation is

$$P = \left(\binom{15}{9} + \binom{15}{9} + \binom{15}{10} + \binom{15}{11} + \binom{15}{12} + \binom{15}{13} + \binom{15}{14} + \binom{15}{15}\right)\left(\frac{1}{2}\right)^{15} \approx 0.3036.$$

Approximately 30% of the time, we would expect to see this division or worse from a population with equal numbers of each.

Finally, students familiar with statistics can create a 95% confidence interval for the proportion of **Af** midges. This is given by $0.6 \pm 2\sqrt{\frac{(0.6)(0.4)}{15}}$.

The data suggest that the true proportion of **Af** midges is in the interval defined by 0.6 ± 0.25. The sample containing 9 **Af** midges and 6 **Apf** midges could have been drawn from a population that contained anywhere from 35% **Af** midges to 85% **Af** midges.

Both the empirical and theoretical arguments tell us not to worry about the difference in numbers. Based on the data, there is no reason to believe there are more of one type than another and, therefore, no reason to weight the solution.

After deciding not to weight the data, this group fit a least-squares line to each data set, as in Solution 1. If the

data are indeed linear, the students knew that the residuals from a least-squares fit should be approximately normally distributed. The residuals are the differences between the actual data and the linear fit, that is, the *errors* in the least-squares fit. This means that close to 68% of the data will fall within one standard deviation of the residuals of the line, and 95% within two standard deviations of the residuals of the line. (**Figure 12**.)

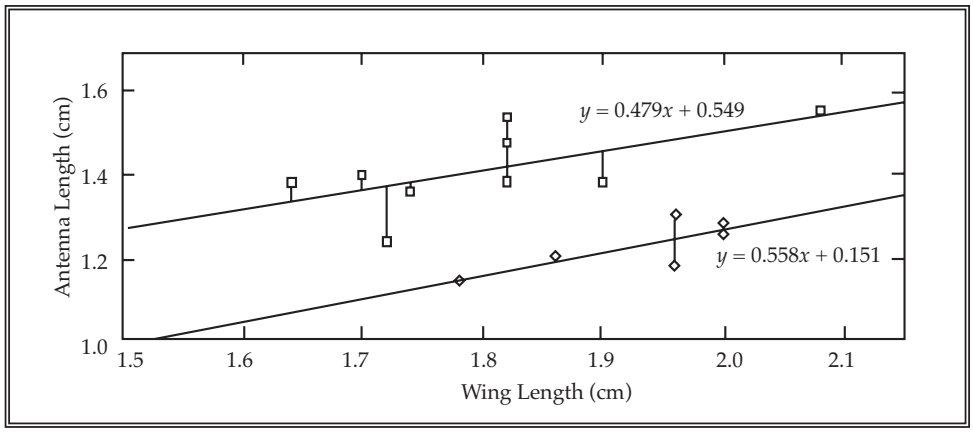


FIGURE 12. RESIDUALS FROM LEAST-SQUARES FIT.

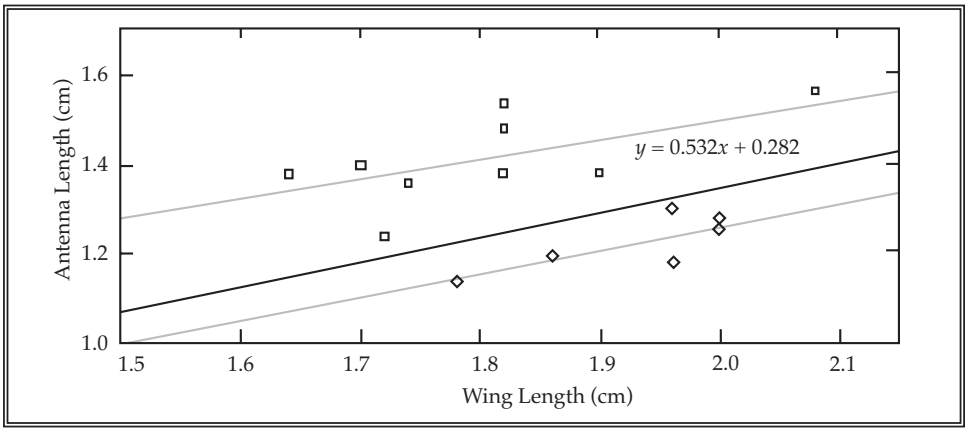


FIGURE 13. EQUI-PROBABLE BOUNDARY AND LINEAR FITS.

The standard deviation of the residuals for the **Apf** linear fit is 0.036, while for the **Af** midges is 0.073. For the **Af** midges, 68% of the data should fall between the lines $y = 0.479x + 0.476$ and $y = 0.479x + 0.622$. A similar boundary exists for the **Apf** midges, $y = 0.558x + 0.115$ and $y = 0.558x + 0.187$. A midge represented by the point that is both one standard deviation from the **Af** line, on $y = 0.558x + 0.187$, and one standard deviation from the **Apf** line, on $y = 0.479x + 0.476$, is just as likely to come from one data set as the other. The midge with a wing length of 3.66 centimeters and an antenna length of 2.23 centimeters is this midge. An equi-probable boundary can be found by equating the one, two, and three standard-deviation lines from each fitted line. This boundary is linear, and has the equation $y = 0.532x + 0.282$. **Figure 13** illustrates the equi-probable boundary along with the two least-squares lines. Any midge below this equi-probable boundary line is considered to be a dangerous **Apf** midge.

As you can see, there were a number of different solutions to the midge problem; some very sophisticated and others very straightforward. In each case, the students had to decide which was most important, being accurate or being safe. Regardless of the approach, students always seemed to enjoy working on the problem and appreciated the many different solutions their classmates created. Naturally, they each thought their own solution was the best. Some mistakes were made in the process, but they all thought a lot about how the mathematics they know could be applied to this new and unique problem. Also, the consistency of the mistakes with the weighted averages brought to light some deficiencies in my instruction. With the AP Statistics course beginning in 1996, questions like this midge problem will play an increasingly important role in high-school mathematics. In fact, Problem 6 from the 2001 AP Statistics exam, in which students must distinguish which graduate students are likely to earn their Ph.D. based on GPA and mean credit hours, shares many features with the Midge Problem. ❑

Reference:

Grogan, William L., Jr. and Willis Wirth. 1981. "A new American genus of predaceous midges related to Palpomyia and Bessia (Diptera: Ceratopogonidae)." *Proceedings of the Biological Society of Washington* 94 (4): 1279-1305.

Helen Compton and Dan Teague are Instructors of Mathematics at the North Carolina School of Science and Mathematics. Both are Presidential Awardees for North Carolina. You may email them at Compton@ncssm.edu and Teague@ncssm.edu.

Everybody's Problems concerns teaching high school mathematics courses with real-world problems, particularly problems that are suitable for students at all levels.