Southern Nevada Regional Professional Development Program
March 2004 — High School Edition

rpdp.ccsd.net

In the January 2004 issue of *Take It to the MAT*, we were presented with data showing the percentage of smokers and lung cancer death rates of ten midwestern states. We made a scatterplot of the data to explore a relationship, and then fit an "eyeball" line to model the data. In this edition, we will use a little more sophisticated, yet simple, technique for finding a line to fit the data.

The method we will discuss is known as the *median-median line*. It requires little computation and is a built in function on many graphing calculators. Here's how it works.

First, sort the data on the "*x*" variable; in this case it is percentage of smokers. (The sorted data are shown in the table at right.)

Next, divide the data into three groups, as evenly and symmetrically as possible. In this case, we have 10 pieces of data, so we will divide them into 3, 4, and 3 observations. (Note: It is important that equal *x* values are placed in the same group. It's not a problem in our situation as both *x* values of 23.3 were already in the middle group. If placing equal *x*-values forces a break from symmetry, we favor placing more observations in the outer groups than the middle group while maintaining group sizes as close to one-third of the data set as possible.)

Now find the median values for *x* and *y* in each of the three groups, and write them as ordered pairs. In the low group, the median point is (21.1, 52.2). Note that we are finding median of *x* and the median of *y* separately—do not find the median of *x* and then use its corresponding *y*-value for the median of *y*. In the middle group, the median point is (22.8, 53.05). In the high group, the median point is (27, 65.4). Plot these three points on the scatterplot in a way that they can be distinguished from the actual data.

Plot a line through the low and high median points. In our case, the (thin red) line through the points (21.1, 52.2) and (27, 65.4) is $y \approx 2.2x + 5.0$, or better yet, *deaths per 100K* $\approx 2.2 \cdot$ *smoker pct* $+ 5.0$.

The last step is to adjust the line for the median point of the middle group. We do this by translating the line toward the middle group's median point by one-third of its vertical distance from the line we just created. The median point of the middle group is (22.8, 53.05) and the line goes through approximately (22.8, 56.00). The vertical distance is about 2.95, so the line will be adjusted downward, toward the middle group median point, by about 0.98. (Note: use unrounded values when doing this computation.) Translating our line down by 0.98 gives us *deaths per 100K* $\approx 2.2 \cdot$ *smoker pct* $+ 4.0$. This is the median-median (heavy blue) line. It is one method to fit a line to a set of data, requires little computation, and is easily accessible for most students.

How does it compare with our "eyeball" line last month?

	State	Smokers_Pct	Deaths_per_100K
1	MN	19.8	46.3
2	KS	21.1	55.7
3	NE	21.4	52.2
4	SD	22	52.8
5	IL	22.3	53.3
6	IA	23.3	58.2
7	ND	23.3	47.3
8	WI	24.1	51.8
9	IN	27	65.4
10	MO	27.2	67.5

(Low: rows 1–3; Middle: rows 4–7; High: rows 8–10)

adjust $\frac{1}{3}$ of the distance