**Mean**

A complete worksheet for using this sketch with students is available as a separate download.

You can think of the six points labeled *x*1 through *x*6 as representing six data values such as six heights, six incomes, or six cholesterol levels. Try dragging the points. Notice that you cannot change their *y*-coordinates. We are only interested in their *x*-coordinates.

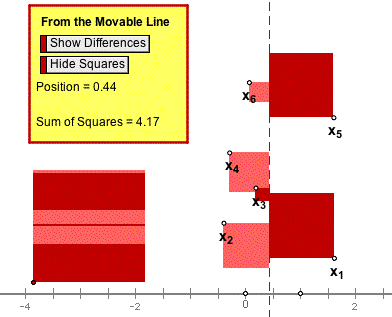
The vertical dashed line is movable. You use it to look for a good place to define as the

"center" of the six points.

Two tools for finding a "center" are shown.

 Press the *Show Differences* button to show the differences between the points and the line and the algebraic sum of these distances. Drag the dotted line and notice how these values change.

 Hide the differences and show the squares. Drag the dotted line and notice how these values change.



**Questions**

1. Describe at least two ways to get the sum of differences to be negative.

2. What can you do to get the sum of differences to be zero?

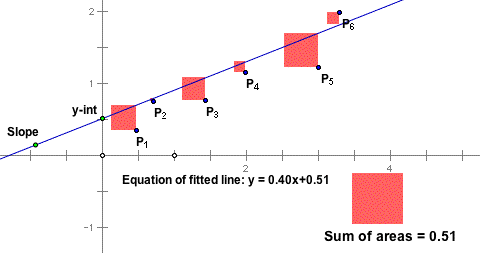
3. If you change the positions of the data points, is it always possible to adjust the dashed line so that the sum of differences is zero? Why or why not?

4. Write a definition for the center of data values that uses the sum of differences.

5. Under what circumstances will the sum of squares be zero?

6. If you change the positions of the data points, is there always a position of the dashed line that makes the sum of squares zero? Why or why not?

7. How can you use the sum of squares to define a center?



Six data points have been placed in the coordinate plane and a line is drawn through them. Move the line by dragging its *y*-intercept or by dragging the point that determines its slope until it appears to be a good fit to the six data points.

From each data point the square of the residual from the point to the line has been constructed. The sum of these squares results in a larger square whose area can be minimized by dragging the line. This then becomes a criteria of best fit; that is, the least squares best fit of a line through the points is that line which minimizes the sum of the squares of the residuals.

**Questions**

1. Press the *Show Regression Line* button to compare your estimate of the best fit with the actual best fit.

2. With both lines hidden, arrange the points so that you think the *r*2 value is:

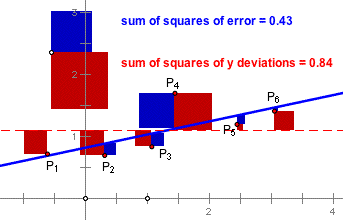
o less than 0.2,

o greater than 0.8,

o between 0.4 and 0.6

Check your estimates by showing the regression line and the correlation coefficient.

3. With the regression line hidden, move the points so that there is an outlier. Attempt to find the best fit regression line and compare your result with the actual. Describe the effect that the outlier has on the least squares line. Why is it that the least squares regression line is said not to be resistant to outliers?



As in the Least Squares visualization, six data points are arranged in the coordinate plane and may be moved about at will. The least squares regression line for the six points is drawn.

Two buttons control hiding and showing of:

 the squares of the deviations of the points from the mean of their *y-*values (the red squares), and

 the squares of the errors or residuals from the points to the regression line (the blue squares).

The sum of the areas of the red squares is shown as a large red square - the total squared deviation that we are trying to explain with a linear model. Likewise, the sum of the areas of the blue squares is shown as a large blue square - the total squared error that remains unexplained by the linear model.

The ratio of these two areas (blue / red) is proportion of the deviation that remains unexplained. One minus this ratio is proportion of the deviation that is explained.

**Questions**

1. Why is it that the big blue square will always have less area than the big red square?

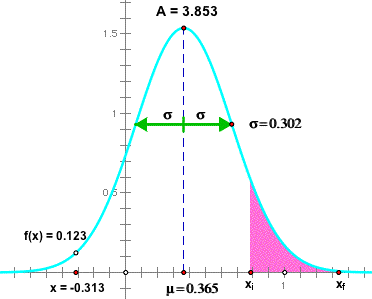
2. Drag point *P*1 along the the regression line until it becomes an outlier. Notice that as you get farther and farther from the other five points, the *r*2 value becomes closer and closer to one. Explain why this is happening and comment on how it can happen that one point can provide a high correlation.

3. Drag the six points until they all have approximately the same *y*-values. Notice that r squared becomes nearly zero. Explain why this has to be so.

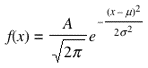
4. Find two other arrangements of the six points that produce a zero *r*2 value and explain why they do so in terms of the red and blue squares.

5. Arrange the six points so that the *r*2 value is one. Explain why your arrangement works.

6. Is there any arrangement of the six points that has an *r*2 value of one but in which the points are not all on a line? Why, or why not?



This sketch allows you to change the parameters of a normal curve and compute the area under any portion of it.



The formula for the normal curve is shown above. There are three parameters to change, each corresponding to a point in the sketch.

 The parameter *A* controls the maximum height of the curve, which is . You can drag the point at the top of the curve to change its height.



 The mean (µ in the formula) is controlled by the point labeled mean in the sketch.

 The standard deviation is σ.

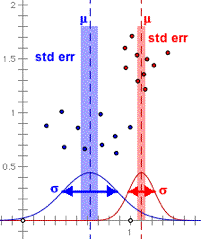
Press *Show Integral* to get two control points, *xi* and *xf*. The area under the curve between these two values will be shown in pink and computed numerically.

**Questions**

1. Demonstrate the assertion that about two-thirds of the area under a normal curve lies within plus or minus one standard deviation of the mean.

2. Use the sketch to find the proportion of the area under the curve that lies to the right of the mean plus one standard deviation.

3. Lengths of tadpoles in a certain pond are found to be approximately normally distributed with a mean of 1.25 inches and a standard deviation of .25 inches. Use the sketch to find the chances of finding a tadpole with a length greater than 1.75 inches.



The red and blue points in this graph represent data from two samples of 10 values from each of two normally distributed populations. (Only the *x*-coordinates of the points are relevant.) You are interested in whether the means of the two distributions are significantly different.

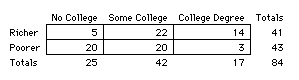
The sketch shows a normal curve for two populations each with a mean and standard deviation corresponding to the mean and standard deviation of the two samples. The standard error (the standard deviation divided by the square root of the number in the sample) for each sample is shown as the half-width of a vertical rectangle. the *t* statistic and *p* value for an unpaired difference of means is shown.

**Questions**

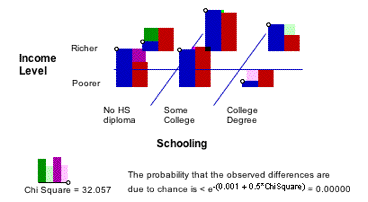
1. Press one of the hide buttons so that just one sample and distribution are shown. Drag one data point back and forth from left to right. Describe the effect on the normal curve of dragging the point when it is close to the mean as compared to when it is far from the mean. What does your observation say about whether the standard deviation is resistant to outlier effects?

2. Select all of the red points, and drag one of the selected points. (This should have the effect of moving all the points at once.) Describe the effect that this has on the mean and standard deviation of the sample. Describe the effect on the *p* value for the difference of means. Explain what you see.

3. What will be the effect on the *p* value of dragging just one point far from the mean of its sample? Predict first! Explain what you observe.



The data displayed here come from a random sample of the 1990 census data for people in Sonoma County, California who are between the ages of 30 and 60. Each of the cells in the table at right show the count of the number of people who fall into that cell; for example, there were five people with no college education who fell into the "Richer" category. Richer means an income of greater than 23000 (the median income for the sample).



You can change the frequency for a cell by dragging the point at the top of the blue bar in that cell. As you drag, the expected values and the chi-square statistic are recomputed. At right, you can see that the chi-square statistic for this situation is 23.7 and that this is high enough that the p-value is very close to zero.

The point of this sketch is that the chi-square algorithm has a geometric visualization. Each cell contributes an amount to the total chi-square corresponding to the width of a rectangle whose area is equal to the square of the difference between the actual and expected values.

**Questions**

1. Is there any way to adjust the cell frequencies so that the chi square = 0? Describe the conditions under which this can be true.

2. Is it possible to have exactly on cell frequency different from its expected value? Why or why not?

3. Is it possible to have both frequencies in one column different than their expected values but all the other four frequencies equal to their expected values? Why or why not?

4. Under what conditions is the chi square very large? (See if you can produce a chi square

> 80.)

5. Suppose two cells have the same difference between observed and expected values, but one has a high expected value and the other has a low expected value. Will they both contribute the same amount to the chi square? Why or why not?