INTRODUCTION TO SMOOTHING

One aspect of regression is to see how the "center" of the conditional distributions varies as a function of the explanatory variable -- e.g., to express \( E(Y|X = x) \) as a function of \( x \).

A **smooth** is a curve constructed to go through or close to all points \((x, E(Y|X = x))\) (a "mean smooth") or through or close to all points \((x, \text{med}(Y|X = x))\) (a "median smooth").

**Example:** In the fish data, we have seen both a median smooth (transparency) and a lowess mean smooth (constructed by arc).

**Note:** The median smooth was easy to construct for the fish data, since there were just a few values of the explanatory variable.

**Example:** In trying to construct a median smooth for the haystack data, we need to choose the number of "slices," introducing the idea of a **smoothing parameter**.

**Note:** 1. What does the haystack smooth help us see in the data? 2. Arc also has a "slide smooth" function illustrating how a parameter in involved in creating a smooth.

The **lowess** (locally weighted scatterplot smoother) **smooth** can be found on most statistical software.

**Outline of how the lowess curve is calculated**

- Start with data points \((x_1, y_1), \ldots (x_n, y_n)\).
- Select a smoothing parameter \( f \) between 0 and 1. (We'll use \( f = 0.5 \) for illustration.)
- For each \( i \),
  - a. Look at the half (if \( f = 1/2 \); 1/4 if \( f = 1/4 \), etc.) of the data with \( x \) values closest to \( x_i \).
  - b. Fit a line (using weighted least squares -- we may talk about this later) to these points in a way that gives more weight to points with \( x \) closest to \( x_i \).
  - c. Replace \( y_i \) with \( y_i' = \) the \( y \)-value of the point on this line corresponding to \( x_i \). (So \( y_i' \) "adjusts" \( y_i \) to be influenced by nearby data points.)
- After doing this separately for each \( i \), repeat the procedure using points \((x, y_i')\) (so the effect of points away from the trend will probably be less.)
- After a few iterations of this process, connect all the current "adjusted" points.