

Class #5 - Wednesday February 17
Paired Data: Scatterplots and the Correlation Coefficient

Textbook readings: Ross, Sec 2.5: Paired Data & Sec 3.7: Correlation Coefficient

Intro: So far we have looked at data sets for a single variable (or “attribute”) for each member of the sample, e.g., x_i = weight of individual i . Now suppose the data set contains values for a second variable for each individual.

Example: Suppose we have a sample of 23 individuals, and we measure each individual’s height and weight. Then we have a **set of paired data**:

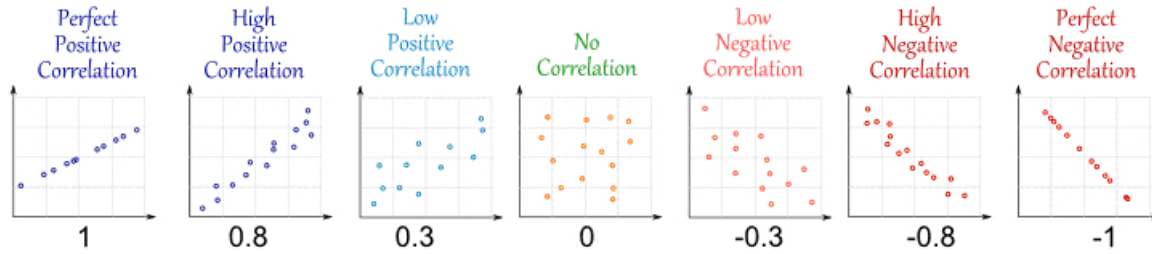
| Individual # | Weight (lbs) | Height (inches) |
|--------------|--------------|-----------------|
| 1 | 132 | 59 |
| 2 | 114 | 58 |
| 3 | 187 | 78 |
| 4 | 214 | 84 |
| 5 | 158 | 71 |
| 6 | 212 | 72 |
| 7 | 175 | 69 |
| 8 | 147 | 59 |
| 9 | 173 | 69 |
| 10 | 182 | 72 |
| 11 | 194 | 78 |
| 12 | 215 | 80 |
| 13 | 180 | 74 |
| 14 | 147 | 68 |
| 15 | 168 | 76 |
| 16 | 157 | 70 |
| 17 | 168 | 73 |
| 18 | 179 | 68 |
| 19 | 216 | 74 |
| 20 | 200 | 77 |
| 21 | 194 | 75 |
| 22 | 147 | 68 |
| 23 | 156 | 67 |

We have discussed methods of looking at the distribution of heights or the distribution of weights *separately*: frequency tables and histograms, mean values, standard deviations. But these wouldn’t tell us anything about the *relationship* between height and weight.

The first thing we can do is visualize the paired data values with a **scatterplot**, where we plot each data point as an ordered pair on an xy -coordinate system.

Exercise: Enter the data above into a spreadsheet and create the scatterplot: (1) Select the two columns of data; (2) use “Insert → Chart ...” (either from the menus or the chart icon in the toolbar); (3) select “Scatter” for chart type; (4) customize the chart title, axis labels, etc.

Additional Examples: See <http://www.gapminder.org/world> and <http://www.nytimes.com/interactive/2014/03/15/business/higher-income-longer-lives.html> for scatterplots showing life expectancy and income data



Correlation Coefficient: For a sample of paired data (x_i, y_i) , the sample correlation coefficient is a statistic that measures the degree to which the variables x and y are correlated, i.e.,

- as x -values increase, do the y -values also tend to increase? (positive correlation)
- as x -values increase, do the y -values tend to decrease? (negative correlation)

Definition: The correlation coefficient r is defined as

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

where s_x and s_y are the standard deviations of the $\{x_i\}$ and $\{y_i\}$, respectively.

- the terms $(x_i - \bar{x})(y_i - \bar{y})$ in the sum in the numerator are positive when the deviations $(x_i - \bar{x})$ & $(y_i - \bar{y})$ have the same sign, i.e., when relatively large x_i are paired with relatively large y_i and small x_i correspond to small y_i (read Ross, Sec 3.7, pp121-123)
- on the other hand, $(x_i - \bar{x})(y_i - \bar{y})$ is negative if the deviations $(x_i - \bar{x})$ & $(y_i - \bar{y})$ have opposite signs, i.e., large x_i are paired with small y_i and small x_i with large y_i
- the denominator serves to “normalize” or “standardize” the value of r , so that r is between -1 and 1
- spreadsheet function: =correl(data_x, data_y)

Exercise: Would you expect height and weight to be positively correlated, negatively correlated, or uncorrelated? What does your scatterplot indicate?

Calculate the correlation coefficient for the example dataset, using the spreadsheet function =correl(data_x, data_y).