

Displaying Data: Paired Numerical Data

- **response variable**
measures a characteristic of interest in a study; the aim is to determine how this variable is affected by variation in some other quantity, namely, an ...
- **explanatory (predictor) variable**
a variable which may turn out to influence the outcome of the response variable
- **scatterplot**
displays paired data (x represents the explanatory variable, y the response variable) as points (x, y) in a coordinate plane

Describing Data: Paired Numerical Data

To investigate the possible relationship between the variables, look for an overall pattern in the plot and be on the watch for any **outliers** or deviations from such patterns

- **association**
tendency for change in one variable to be accompanied by change in the other
- **direction**
variables display a **positive association** if larger values of one tend to be paired with larger values of the other, and a **negative association** if larger values of one tend to be paired with smaller values of the other
- **form**
shape of the plot, including **clusters** of data points; **linear** relationships are most important
- **strength**
how closely the points conform to the overall shape of the plot

Summarizing Data: Paired Quantitative Data

- **correlation** (r)

a measure of the direction and strength of *linear* relationship between quantitative variables; where \bar{x} and s_x represent the mean and standard deviation of the explanatory variable data and \bar{y} and s_y the similar quantities for the response variable data, then, if (x, y) is any one of the data points, we compute the corresponding standardized values

$$z_x = \frac{x - \bar{x}}{s_x}, \quad z_y = \frac{y - \bar{y}}{s_y}$$

whence the correlation is computed as the average of the products of the standardized values:

$$r = \frac{\sum_{(x,y)} z_x z_y}{n - 1}$$

(Note again the division by $n - 1$. Also, the summation symbol $\sum_{(x,y)}$ indicates that the products

of the standardized values are to be summed over all the points in the data set.)

Analyzing Data: Paired Quantitative Data

Interpreting the value of r :

- positive values of r indicate a positive association between x and y ; negative values of r indicate a negative association between x and y
- r always lies between -1 and 1 , with values close to 0 indicating weak association, values close to 1 a strong positive association, and values close to -1 a strong negative association

Note that

- correlation can only be computed for paired quantitative variables and *does not apply to categorical variables*
- while it is possible to compute the correlation for any pair of quantitative variables, *only linear associations* are evaluated by the value of r
- r is highly sensitive to outliers, so *the presence of an outlier can dramatically alter the value of the correlation* of a paired data set; study correlation values with and without outliers to evaluate this effect
- there may be a strong association between variables without there being a cause/effect relation between them: *association does not signify causation*. Sometimes, there is a third **lurking variable** that stands behind the other two variables as a common determining factor.