Correlation

- **Odds-ratio**: a measure of the association between two categorical variables (possibly adjusting for a third variable)

- **Pearson's Correlation** coefficient: a measure of the linear association between two continuous variables

- **Spearman's Rank Correlation**: measure of association between two ranked variables (applies to continuous or ordinal data)
Correlation

Population coefficient between $X$ and $Y$ is the average of the product of standardized $X$ and $Y$ values

$$\rho = \text{Average} \left( \frac{(X - \mu_X)}{\sigma_X} \cdot \frac{(Y - \mu_Y)}{\sigma_Y} \right)$$

Pearson Correlation

$$r = \frac{1}{n-1} \sum \frac{(X_i - \bar{X})}{s_x} \cdot \frac{(Y_i - \bar{Y})}{s_y}$$

Spearman's Rank Correlation uses the ranks of $X$ and $Y$ in place of the raw data in the above formula.
Values

• **Correlations are between -1 and +1**
  
  − -1: perfect negative linear association
  
  − +1: perfect positive linear association
  
  − 0: no association
    
    − this does not imply X and Y are independent
    
    − no linear association; there may be a nonlinear relationship!

• **Correlations can be misleading if there are**
  
  − outliers

  − nonlinear relationships
Scatterplots

Always plot the data!

- often can visually determine whether a relationship exists between two variables in a scatter plot
- scatterplot can show outliers (these can greatly impact the correlation coefficients)
- scatterplot can show nonlinear relationships
  - correlation is a measure of linear association
  - may have a perfect nonlinear relationship, but no association as measured by the correlation coefficient!

• If all is OK, correlation provides a numerical summary of the linear association between two variables
Lung Cancer Deaths

Positive association: the higher the cigarette consumption in a country in 1930, the higher the death rate from lung cancer in 1950

correlation = 0.74

Ecological Correlation: based on averages

Cannot conclude from this data that death rates for lung cancer are higher for those individuals who smoke more!
Correlation between X and Y?
Outliers

- Correlation?
Correlation between X and Y?
Summary

• Correlation coefficients are dimensionless numbers and are unchanged by
  – adding a constant to one of the variables
  – multiplying one of the variables by a positive constant

• Correlation can be misleading if there are outliers or a nonlinear relationship

• Correlation measures linear association, not necessarily causation; the observed association may be due to a third variable that simultaneously influences both variables - confounding.

• Ecological correlations are correlations based on rates or averages, and tend to overstate the strength of association for individuals